

格致方法·定量研究系列 吴晓刚 主编

事件史和生存分析（第二版）

[美] 保罗·D.埃里森（Paul D. Allison）著
范新光 译 刘孟宇 校

- 革新研究理念
- 丰富研究工具
- 最权威、最前沿的定量研究方法指南

62

格致出版社 上海人民出版社



本书通过严谨的统计语言和生动的例子，详实而系统地介绍了处理事件史数据的方法。作者特别关注了回归方法，即事件的发生依赖于一个或多个解释变量。他解释了构成事件史分析基础的统计模型，介绍了在实际分析中如何进行操作，包括数据管理、成本和一些有用的计算机软件。对于希望了解事件史数据处理方法的读者而言，本书不失为一本全面扼要的手册。

主要特点

- 系统严谨，直观扼要
- 本书作者长期从事事件史分析的创新和应用分析，提供了极具实践指导意义的方法指引
- 事件史数据是社会科学极为重要的数据类别，本书涵盖了事件史分析的主要方法

上架建议：社会研究方法

ISBN 978-7-5432-2096-6



您可以通过如下方式联系到我们：
邮箱：hibooks@hibooks.cn


微信


天猫

定价：32.00元
易文网：www.ewen.co
格致网：www.hibooks.cn

格致方法·定量研究系列 吴晓刚 主编

事件史和生存分析(第二版)

[美]保罗·D.埃里森(Paul D. Allison)著
范新光 译 刘孟宇 校

SAGE Publications, Inc.

格致出版社 上海人民出版社

图书在版编目(CIP)数据

事件史和生存分析:第二版/(美)保罗·D.埃里森(Paul D.Allison)著;范新光译.—上海:格致出版社:上海人民出版社,2017.6
(格致方法·定量研究系列)
ISBN 978-7-5432-2096-6

Ⅰ.①事… Ⅱ.①保… ②范… Ⅲ.①统计分析-研究 Ⅳ.①C812

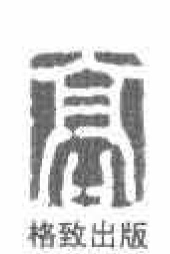
中国版本图书馆 CIP 数据核字(2017)第 103351 号

责任编辑 贺俊逸

格致方法·定量研究系列

事件史和生存分析(第二版)

[美]保罗·D.埃里森 著
范新光 译 刘孟宇 校

出 版	世纪出版股份有限公司 格致出版社 世纪出版集团 上海人民出版社 (200001 上海福建中路 193 号 www.ewen.co)  编辑部热线 021-63914988 市场部热线 021-63914081 www.hibooks.cn	印 刷	浙江临安曙光印务有限公司
		开 本	920×1168 1/32
		印 张	5.75
		字 数	94,000
		版 次	2017 年 6 月第 1 版
发 行	上海世纪出版股份有限公司发行中心	印 次	2017 年 6 月第 1 次印刷

ISBN 978-7-5432-2096-6/C·179 定价:32.00 元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法·定量研究系列”丛书，精选了世界著名的 SAGE 出版社定量社会科学研究丛书，翻译成中文，起初集结成八册，于 2011 年出版。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的热烈欢迎。为了给广大读者提供更多的方便和选择，该丛书经过修订和校正，于 2012 年以单行本的形式再次出版发行，共 37 本。我们衷心感谢广大读者的支持和建议。

随着与 SAGE 出版社合作的进一步深化，我们又从丛书中精选了三十多个品种，译成中文，以飨读者。丛书新增品种涵盖了更多的定量研究方法。我们希望本丛书单行本的继续出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总 序

2003年，我赴港工作，在香港科技大学社会科学部教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课"社会科学里的统计学"(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课"社会科学中的定量分析"为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的"消费者"而非"生产者"。作为"消费者"，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少

量重复,但各有侧重。“社会科学里的统计学”从介绍最基本的社会研究方法论和统计学原理开始,到多元线性回归模型结束,内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型,以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法,将重点放在因变量为定类数据的分析模型上,包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型,包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生,一直鼓励和支持我将两门课的讲稿结集出版,并帮助我将原来的英文课程讲稿译成了中文。但是,由于种种原因,这两本书拖了多年还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩,每本书都写得通俗易懂,与我的教学理念是相通的。当格致出版社向我提出从这套丛书中精选一批翻译,以飨中文读者时,我非常支持这个想法,因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力,还要有对实质内容有较深的理解能力,而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容,只有语言能力是远远不能胜任的。在短短的一年时间里,我们组织了来自中国内地及香港、台湾地区的二十几位

研究生参与了这项工程，他们当时大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾"中央"研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦；哈佛大学社会学系博士研究生郭茂灿和周韵。

参与这项工作的许多译者目前都已经毕业，大多成为中国内地以及香港、台湾等地区高校和研究机构定量社会科学方法教学和研究的骨干。不少译者反映，翻译工作本身也是他们学习相关定量方法的有效途径。鉴于此，当格致出版社和 SAGE 出版社决定在"格致方法·定量研究系列"丛书中推出另外一批新品种时，香港科技大学社会科学部的研究生仍然是主要力量。特别值得一提的是，香港科技大学应用社会经济研究中心与上海大学社会学院自 2012 年夏季开始，在上海(夏季)和广州南沙(冬季)联合举办《应用社会科学研究方法研修班》，至今已经成功举办三届。研修课程设计体现"化整为零、循序渐进、中文教学、学以致用"的方针，吸引了一大批有志于从事定量社会科学研究的博士生和青年学者。他们中的不少人也参与了翻译和校对的工作。他们在

繁忙的学习和研究之余，历经近两年的时间，完成了三十多本新书的翻译任务，使得“格致方法·定量研究系列”丛书更加丰富和完善。他们是：东南大学社会学系副教授洪岩璧，香港科技大学社会科学部博士研究生贺光烨、李忠路、王佳、王彦蓉、许多多，硕士研究生范新光、缪佳、武玲蔚、臧晓露、曾东林，原硕士研究生李兰，密歇根大学社会学系博士研究生王骁，纽约大学社会学系博士研究生温芳琪，牛津大学社会学系研究生周穆之，上海大学社会学院博士研究生陈伟等。

陈伟、范新光、贺光烨、洪岩璧、李忠路、缪佳、王佳、武玲蔚、许多多、曾东林、周穆之，以及香港科技大学社会科学部硕士研究生陈佳莹，上海大学社会学院硕士研究生梁海祥还协助主编做了大量的审校工作。格致出版社编辑高璇不遗余力地推动本丛书的继续出版，并且在这个过程中表现出极大的耐心和高度的专业精神。对他们付出的劳动，我在此致以诚挚的谢意。当然，每本书因本身内容和译者的行文风格有所差异，校对未免挂一漏万，术语的标准译法方面还有很大的改进空间。我们欢迎广大读者提出建设性的批评和建议，以便再版时修订。

我们希望本丛书的持续出版，能为进一步提升国内社会科学定量教学和研究水平作出一点贡献。

吴晓刚

于香港九龙清水湾

序

社会科学家分析许多感兴趣的现象时，关注的是事件的时间性：生命期望、在失业之后重新找到工作所需要的时间、婚姻存续的时长、累犯的间隔时间，等等。几乎所有关于事件时间的数据的一个关键特征是删截(censoring)：例如，在一个关于累犯的研究中，研究者想记录犯人从监狱里被释放出来的一年内再次被捕的情况。尽管某些人最后有可能在接下来直到研究结束的时期内再次被捕，但也有一些初犯在这一时间段里并没有再次被捕。

研究此类事件发生时间的方法在许多学科领域都有进展，包括社会学中的事件史分析、工程领域的失效分析以及更广泛意义上的生物统计领域的生存分析。这些术语反映了不同学科的关注，但它们在本质上是相通的。一旦认识到它们基本的研究单位，我们便发现了研究事件时间性的一种共同方法。而这一方法就

是本书的主题。

有很多术语可以去称呼它,生存分析在社会科学领域里使用得最为广泛。在保罗·埃里森第二版重新命名的关于生存分析的小册子中,他向我们展示了一个关于这一主题涵盖甚广的介绍,同时将文笔集中在生存回归模型上,这一模型将事件发生时间和解释变量联系到了一起。生存回归模型——更深入地说,Cox比例机会模型——已经被应用于从生存数据得出因果推断的研究之中,这在社会科学中变得更为普遍,同时也应用到了对基于特殊设计的预测研究之中,例如对信用卡拖欠的研究中。

尽管埃里森教授阐述了很多种生存分析的方法,而且其中的一些方法本身还很复杂,但他对生存分析的解释是清楚明了与通俗易懂相结合的典范。本书的突出之处在于,它讨论了离散时间数据的方法,这在以往的讨论中往往被忽略;并且,它不仅关注单次独特事件(典型的例子如死亡),而且也关注了多重事件(包括"竞争性"事件,例如婚姻终结是由离婚或者死亡所导致),以及多次发生事件(例如失业的重复发生)。

埃里森教授这本书的第一版有一大批忠实的读者。我相信本书的第二版将同样会对新一代把生存分析应用到他们研究之中的社会科学家们有所裨益。

约翰·福克斯

第二版前言

两版间隔的 30 年是很长的一段时间。第一版(出版于 1984 年)已经有些过时,尤其是对于电脑软件和程序命令的描述。但是上一版的基本结构和大部分内容依然得到保留。我采用了一种不一样的策略:首先介绍的是简单地通过 logistic 回归就可以执行的离散时间方法,然后转向参数连续时间方法,紧接着是半参数 Cox 回归。

在新版中,最明显的变化是标题,从《事件史分析:纵贯数据的回归》(*Event History Analysis: Regression for Longitudinal Event Data*)到简单的《事件史和生存分析》(*Event History and Survival Analysis*)。尽管"事件史分析"这一术语出色地将这些方法广泛地应用到了所有类型的事件数据之上,但大多数研究者如今将其称为生存分析,这是因为他们的生物统计学背景使他们一直关注死亡事件发生的建模。

大部分数据集也得到了更新，主要因为原始数据在过去 30 年里已经丢失了。新的数据集可以在 http://www.statisticalhorizons.com/resources/data-sets 上下载，包括了 Stata 和 SAS 数据集。

以下是每章的主要变动和增补的概述：

第 2 章（“离散时间方法”）现在提供了关于删截更为详细的讨论，包括了如何对信息删截进行敏感性检验的例子。

第 3 章（“连续时间数据的参数法”）对加速失效时间模型给予了更多的关注。同时，本章更加强调了对结果的解释和评估模型拟合度的方法。

第 4 章（“Cox 回归”）借助于程序陈述法和分割区间法，对时变解释变量进行了更加详细的介绍。同时，本书包含了对检验比例机会假定和处理连接数据（tied data）方法的更具扩展性的讨论。最后，有一个简短的部分介绍了如何用 Cox 模型做预测。

第 5 章（“多种类事件”）现在增加了检验多种类型事件的系数差别的例子。这里也增加了累计发生方程的新版块。这一方程是处理竞争性风险的一种替代性且越来越流行的方法。

第 6 章（“重复事件”）描述了一些在 1984 年还没有出现的新方法，包括处理事件计次数据的负二项模型、稳健标准误和共享异质性（随机效应）模型。基于间隔次数的方法被从基于发生次数的方法中区分开来。

所有例子的计算命令(使用 SAS 和 Stata)现在可以通过网上获得(http://www.statisticalhorizons.con/resoures.books 和 www.sagepub.com/allisonevent)。我会尽可能地更新它们。

目 录

第1章

导　言

社会科学的几乎所有领域对事件及其产生的原因都有极大的兴趣。犯罪学家研究犯罪、逮捕、定罪和监禁;医学社会学家关注住院治疗、拜访医生和精神病发作;在工作和职业研究领域,职业变迁、晋升、下岗和退休得到了充分的关注;政治学者对骚乱、革命以及政府的和平演变保持着兴趣;人口学家则关注出生、死亡、结婚、离婚和迁徙。

在上述的每一个例子中,一个事件(event)由一段特定时期内的某些质变构成。通常而言,"事件"这一术语并非依靠一些量变项去描述一个渐变的现象,而是表示"在此之前"与"在此之后"之间所发生的相对剧烈的变化。

由于事件是在时间变迁的语境下被定义的,所以通过收集事件史数据去研究事件及其原因的方法越来越被学界所接受。这一数据最基本的形式可以描述为,事件史是对事件在个体样本或总体样本中何时发生的纵贯记录。例如,一个调查会向受访者询问他们

的结婚日期(如果有的话)。如果目的是要研究事件的原因,那么事件史也要包括可能的解释变量的数据。这其中包括不随时间变动的变量(如种族),也包括可能随时间变动的变量(如收入)。

尽管对于研究事件的起因而言,事件史看似完美,但仍然存在两个给标准统计方法(比如线性回归)带来麻烦的特征——删截和时变解释变量。事实上,应用标准方法会带来严重的偏误或信息缺失。然而,在过去40年,已经出现了几种包容事件史数据这两个特征的创新方法。因此,并不存在事件史分析的单一方法,而是存在一个既有差异又相互补充的相关方法的集合。

带着认为这些方法对各类数据分析和社会科学典型问题颇有帮助的视角,本书将审视它们。值得强调的是,事件的发生依赖于一个或多个解释变量的回归方法将会被重点讨论。尽管本书会对构成事件史分析基础的统计模型给予大量的关注,但是例如数据管理、成本和计算机软件的可行性等实际问题同样会被讨论。在转向这些方法之前,我们首先要讨论将传统方法应用于事件史分析时所面临的问题。

第1节 | 事件史分析的难题

要领会经典方法用于事件史数据分析时的局限性,一个具体的例子显得颇有必要。罗西、伯克和勒尼汉(Rossi, Berk & Lenihan, 1980)对马里兰州立监狱432个犯人在被释放之后进行了一年的跟踪调查,对其累犯现象进行研究。研究感兴趣的事件是逮捕,目的是确定逮捕的发生在多大程度上依赖于某些解释变量。

尽管每次被逮捕的时间是已知的,但是罗西等人仅仅构造了一个虚拟变量(1, 0)来标识个体是否在12个月的观测区间内被逮捕。这个虚拟变量作为一个线性回归的因变量,解释变量包括被释放时的年龄、种族、教育和过往工作经历。尽管这是一个合理的研究方式,但远不理想。除了众所周知的将普通最小二乘法用于虚拟因变量的问题(Long, 1997),将因变量二分化显得武断并且浪费了信息。这一方法之所以武断是因为除了研究在12个月那一点结束之外,分界线并无其他特别的意义。利用相同的数据,研究者同样可

以比较在 6 个月这一分界点之前被逮捕的个体或之后被逮捕的个体。这个方法同样浪费了信息，因为它忽略了分界线两侧的变化。比如，我们可能会怀疑一个刚被释放就被再次逮捕的人要比释放 11 个月之后被再次逮捕的人更倾向于犯罪。

为了避免这些难题，把被释放到首次逮捕的时间间隔作为一个线性回归的因变量显得很有诱惑力。但这又会带来新的问题。首先，在一年的观察区间里，因变量的数值对于那些没有被再次逮捕的个体而言是不可知的或"被删截的"。如果被删截的个案很少，那么将这些样本简单地排除是可接受的。但是在一个样本中，74％的个案被删除，这意味着这些个案被删除会产生很大的偏误（Sørensen，1977；Tuma & Hannan，1978）。一个可选的解决方法是指派观测时间的最大值——在这个例子中是一年——作为被删截的因变量的取值。但这明显会低估真实值并使大量的偏误再次出现。

即使没有一个观测值被删截，研究者依然会面临另一个问题：如何将在观察期间数值变化的解释变量纳入进来？例如，在这个研究中，在被释放后的一年时间里，研究对象在每个月都会被调查以获得他们的收入、婚姻状况、就业状况和其他的变化。尽管稍显笨拙，但将 12 个月中每个月的收入信息都纳入到回归模型当中似乎是合理的。对一个在第 12 个月才被再次

逮捕的人而言,这可能很合理,但它显然不适用于那些被释放之后的第一个月就被逮捕的人——他第一个月后的收入和分析并不相关。的确,在剩下的日子里,这个人一直被关在监狱里,所以他的收入成为了累犯的结果而非原因。简言之,并没有一种令人满意的方法能够将时变的解释变量纳入到线性回归之中去预测一个事件发生的时间。

删截和时变解释变量是事件史数据的两个颇为典型的问题。相对而言,删截是更为普遍的难题,因为解释变量往往只被测量一次。不过,对变量进行定期重复测量的纵贯数据集已经越来越常见。对于多数事件而言,这类数据能够更为准确地捕捉到时变变量的作用。

第 2 节 | 事件史方法综述

毫无疑问，事件史数据对于社会科学而言并不是独一无二的，许多高级的方法在其他学科里已经出现。这也是许多初学者的困惑之源，因为相似的甚至有时相同的想法往往通过非常不同的方式——尤其是社会科学家感到陌生的文本——表达出来。因此，以对这些迅速发展的方法进行简单的历史比较综述作为开场显得很有必要。

从人口学那里，我们获知了最早、最著名的而且仍在被广泛使用的事件史分析方法——生命表。然而，生命表并不属于本书讨论的范畴，一方面因为在经典的人口学文献中它已经被充分地讨论过了（例如，Preston, Heuveline & Guillot, 2000），另一方面因为它并没有涉及包含解释变量的回归模型。但值得注意的是，最有影响力的回归模型之一——考克斯（Cox, 1972）的部分似然方法——正是从生命表背后的基本思想获得了灵感。

尽管生命表从 18 世纪就已经应用于研究之中，但

是直到20世纪50年代末60年代初才出现更为新颖的事件史分析方法。在生物医学领域,迫切需要这些方法的根本原因在于生存数据分析,而且许多事件史分析方法的文献确实是以生存分析的名义出现的。例如,一个实验可能会这样进行:实验动物被注入有毒物或缓和药物,实验者观测在每种治疗方案下这些动物生存的时间。因此,这里的事件是指动物的死亡。由于实验往往在所有动物都死亡之前就已经终止,因此出现了删截。生物统计学家已经发表了大量的文献以寻找分析这些数据最有效的方法(关于文献目录,参见Klein & Moeschberger, 2010)。这些方法已经成为癌症患者生存数据分析的标准步骤。

同时,工程师在分析机器和电子零件故障的数据时面临着相似的问题。他们发展的这些方法——通常被称为"可靠性"分析或"故障时间"分析——在理念上与生物统计学家所发展的方法极为相似但在方向上又存在些许差异(Nelson,2004)。

社会科学家对这一领域的关注有点晚,而且很多年来他们并不清楚其在生物统计学和工程学中的进展。不过,一种将马尔科夫过程(Markov processes)理论应用到社会科学中的富有活力的趋势在20世纪60年代末和70年代初出现了(参见Singer & Spoilerman, 1976)。这一趋向的转折点是图玛(Tuma, 1976)将解释变量引入到连续时间的马尔科夫模型之中。这一革新有效地

弥补了社会学取向和在生物统计学和工程领域所取得的进展之间的鸿沟。经济学家也对该领域作出了重要的贡献(例如,Lancaster, 1992)。

在这一章的剩下部分,我试着描绘出区分事件史数据分析取向的几个主要维度。在有些情形下,有些维度有效地区分了生物统计学、工程学和社会学各自开发的方法;而另一些维度超越了学科的界限。所有这些维度构成了本书余下部分的组织基础。

分布法 vs.回归法

许多关于事件史分析的早期文献可以被视为对事件发生时间或事件发生间隔时间分布的研究。这是诸如生命表分析的主要任务。相似地,马尔科夫过程应用于社会科学现象的分析主要关注的是个体在不同情形下的分布状况。最近上述三个学科传统都将注意力转向回归模型,认为一个事件的发生取决于带有一组解释变量的线性函数。如前文所述,本书将专注于回归模型。

重复事件 vs.非重复事件

考虑到死亡是生物学家最关心的事件,那么生物统计学方法对单一的、非重复事件研究的侧重就不足

为奇了。另一方面,社会科学家关注诸如工作变迁、婚姻状态等可能在个体生命历程中发生多次变动的事件,显然这也是本书重点讨论重复事件的原因。此外需要指出,重复事件的模型往往更为复杂并且带来了许多统计难题,进而对单一事件模型的精通是理解更为复杂的模型的基础。因此,我们需要花费大量的时间在更为简单的非重复事件的案例上。

单一事件 vs.多种类事件

在很多情形下,将所有事件视为无差异的在分析中会显得非常方便。因此,对工作终止的研究并不会区分其中的差异。生命表就会认为所有的死亡是一样的。但是在其他情形下,我们需要区分不同类别的事件。在对工作终止的研究中,区分自愿终止和非自愿终止工作显得很有必要。而在一个研究癌症治疗有效性的研究中,区分癌症导致的死亡和其他原因导致的死亡显得尤为重要。为了包容不同的事件,生物统计学家发展出被称为"竞争风险"(competing risks)的方法,而人口学家提出了多项减缩生命表法。图玛和格罗内维尔德(Tuma & Groeneveld, 1979)提出的一般化马尔科夫模型同样考虑到了多种类事件。然而,再次考虑到多种类事件的引入会使模型更为复杂,最好把它放在后面考虑,直到我们很好地理解了单一事件的模型。

参数法 vs.非参数法

生物统计学家喜欢很少包含事件发生时间分布假设的非参数法。相反，工程师和社会科学家着迷于那些假设事件发生时间（或者事件发生的间隔时间）来自不同类别分布的模型，最常用的是指数分布、Weibull分布和Gompertz分布。沟通这两个取向的主要桥梁是Cox（1972）比例机会模型，它被视为半参数或部分参数的。它被视为参数模型是因为这一模型赋予了回归模型一个指定的函数；而它又可被视为非参数模型是因为它并没有指定事件发生时间分布的准确形式。在这一意义上，它大致类似于没有限定误差项分布形式的线性模型。

离散时间 vs.连续时间

假设事件发生时间被准确观测的方法被称为连续时间方法。在实际操作中，时间往往以离散的单位（尽管单位很小）被测量。当这些单位足够小时，我们通常认为时间是在连续的范围上被测量的。但是，当时间单位很大时（比如月、年或十年）使用离散时间法更为合理（也被称作分组数据法）。尽管连续时间分析的方法占据了事件史文献的主流，但是仍然有相当比例的

文献采用了离散时间法,尤其是在生物统计领域(Brown, 1975; Prentice & Gloeckler, 1978; Mantel & Hankey, 1978; Holford, 1980; Laird & Olivier, 1981)。因为离散时间法更容易被理解和使用,它们可以作为事件史分析基本原理的有益介绍。

第 3 节 | 计算

大部分主要的统计软件包都包含了进行生存分析的命令和步骤。在本书中,我使用了 SAS(9.3)和 Stata(12)进行表中的分析。这是我经常使用的两个软件包,它们在进行生存分析时表现出色。所有分析的计算命令可以在 http://www.statisticalhorizons.com/resources/books 中获取。所使用的数据集也可以在 www.statisticalhorizons.com/resources/data-sets 中获取。

第2章

离散时间方法

这一章介绍单一非重复事件的离散时间方法。尽管这是最简单的情形,但是它包含了许多对于更复杂形式的数据而言的基本思想。同时,这个方法极为实用,能够被应用在众多情境中。它也可以被推广到多种类重复事件的研究中(Allison, 1982)。

第 1 节 | 一个离散时间的例子

让我们从一个实际例子开始讨论。这个样本包括 301 个男性生物化学家，他们都是在 20 世纪 50 年代末和 60 年代初获得了博士学位并且在美国大学研究学院获得了助理教授的教职。对这个样本更为详细的介绍可以参见朗、埃里森和麦金尼斯(Long, Allison & McGinnis, 1979)。从获得助理教授职位的第一年开始，他们被最长观测了 10 年。兴趣事件是他们升任副教授。大部分这类升迁毫无疑问意味着终身教职，但是我们并不能肯定。一些大学并没有在聘任副教授的同时授予终身教职。

由于我们只知道升迁的年份而非准确的月、日，所以这些事件是通过离散时间记录的。表 2.1 显示了在 10 年中每一年拿到副教授职位的生物化学家的人数。其中 217 人得到了升迁，84 个被删截。删截的原因有两个:25 人因为 10 年之后仍然没有拿到终身教职，另外59 人因为在 10 年结束之前离开了学术领域。如表 2.1 所示，这些事件在整个观察时间内每一年的发生

次数也是不一样的。

我们的目的是要构建并估计一个回归模型，在这一模型中，一年时间内升迁的条件概率取决于几个解释变量。其中，三个不随时间改变的变量：*undgrad*，测量他们本科学校的录取竞争性；*phdprest*，测量他们获得博士学位院系的声望；*phdmed* 是一个指示(虚拟)变量，测量他们的博士学位是从医学院而非农学院获得。其他三个变量可能会随时间发生改变：*jobpres*，他们目前任教院系的学术声望；*arts*，发表论文的累计数量；*cits*，每个人发表论文被别人引用的次数。

表 2.1 升迁的年份分布(301 位生物化学家)

年份	升迁数量	删截数量	面临风险的数量	估计的机会
1	1	1	301	0.003
2	1	6	299	0.003
3	17	12	292	0.058
4	42	10	263	0.160
5	53	9	211	0.251
6	46	7	149	0.309
7	31	6	96	0.322
8	15	2	59	0.254
9	7	6	42	0.167
10	4	25	29	0.138
总计	217	84	1 741	0.125

第 2 节 | 离散时间机会

现在我们进行模型的构建。事件史分析的一个核心概念是风险集(risk set)。它是指在每一时间点面临某事件发生风险的个体集合。对生物化学家的样本来说,所有 301 个人在第 1 年都面临升职的风险,因此他们构成了那一年的风险集。在那一年,事实上只有一人被提拔为副教授,因而这个人在第 2 年不再面临风险。有一人在第 1 年的结束离开了学术界,因此这个人也不再属于风险集。结果,第 2 年风险集的数量降低到了 299。在每一年年末,风险集要减去当年事件发生的数量和到年末删截的数量。例如,在表 2.1 中,我们能够看到风险集的数量从第 1 年的 301 减少到了第 10 年的 29。

第二个关键概念是机会率(hazard rate),有时被简化称为机会(hazard)或比率(rate)。在离散时间的限定下,假定个体在给定时间内面临着风险,机会率是指事件在特定时间发生在特定个体身上的可能性。在当前的例子中,机会是指在特定年份里还没有升职的个

体得到提升的可能性。有必要指出的是,机会是一个不可被观测的变量,但是它控制了事件的发生以及其发生时间。它构成了事件史模型的基本因变量。

如果假设机会随年份改变但对于每一年的个体而言都是相等的,研究者可以很容易地得出机会的估计值:在每一年里,用事件发生数除以在风险中个体的数量。例如,在第 3 年,292 个处于风险集中的生物化学家中有 17 个升了职。那么估计的机会是 17/293 = 0.058。对其他年份的估计参见表 2.1 的最后一列。我们看到升职的机会一直到第 7 年都在迅速升高(达到了 0.322),但在接下来的年份里出现了下降。需要注意的是,由于风险集一直处于下降的趋势,即便是升迁的个体数量下降,机会仍有可能上升。例如,第 6 年估计的机会比第 5 年的机会要高,尽管在第 5 年有更多的人升迁。

第3节｜logistic 回归模型

下一步要确定机会如何依赖于解释变量。我们将机会定义为 $P(t)$，即个体在时间 t 发生事件的可能性，这里的前提是个体还没有发生事件。简便起见，我们假定有两个解释变量：x_1 不随时间变化，$x_2(t)$ 在 t 的每个点有不同的取值。以生物化学家为例，x_1 可以是获得博士学位院系的声望，而 $x_2(t)$ 是在年份 t 发表的论文累计数量。

如前所述，我们将 $P(t)$ 写成这些解释变量的线性函数：

$$P(t)=b_0+b_1x_1+b_2x_2(t) \qquad [2.1]$$

其中 $t=1, \cdots, 10$。这里存在的一个问题在于 $P(t)$。因为它表示的是概率，所以取值在 0 到 1 之间；而函数的右边取值没有这一限制。这一模型会产生不可能的预测，并且给计算和解释带来困难。这一问题可以通过对 $P(t)$ 进行 logit 转换得到避免：

$$\log\left(\frac{P(t)}{1-P(t)}\right)=b_0+b_1x_1+b_2x_2(t) \qquad [2.2]$$

由于 $P(t)$是在 0 到 1 之间取值,那么函数的左边取值转变为从负无穷到正无穷。当然,还可以有其他方式的变换,但是 logit 转化是最普遍同时在计算上是最方便的(Long, 1997)。系数 b_1 和 b_2 给出了分别对于 x_1 和 x_2 增加 1 个单位,logit 模型因变量(对数发生比)发生的变化。

方程 2.2 中的模型仍然有一定的局限性,因为它隐含了随时间变化的机会的改变仅仅来源于 x_2 这一时变解释变量的变化。在大多数情形下,我们有理由怀疑机会随时间而自然地变化。例如对于学术升迁,大多数研究机构有清晰的升迁标准,一般而言升迁的截止期限是 6 年。我们能够在表 2.1 中清楚地看到。由于在第 7 年之后机会迅速下降,因此我们允许模型通过纳入时间和时间的平方两个变量来显示出机会的曲线变化。

$$\log\left(\frac{P(t)}{1-P(t)}\right)=b_0+b_1x_1+b_2x_2(t)+b_3t+b_4t^2 \qquad [2.3]$$

如果时间点的数量很少,那么我们可以通过将时间点转化为一系列的分类虚拟变量来允许机会随时间发生的任何变动。

第4节 | 模型估计

下一个问题是估计参数 b_0、b_1、b_2、b_3 和 b_4 的取值。正如我们要考虑的模型，估计最好是通过极大似然法或相近的过程来完成。极大似然法的原理是寻找一组相关系数估计值，使已经被观测到的数据的"可能性"最大化。为了完成这一步骤，研究者首先需要将观测数据的可能性表达为一个包含未知系数的函数，然后运用相关的计算方法使这个函数最大化。这两个步骤尽管在数学上较为困难，但对如何估计一个模型而言都不属于重要的知识。对于更详细的内容，有兴趣的读者可以参看埃里森(Allison, 1982)。幸运的是，这一估计会简化成那些从事二分因变量分析的学者颇为熟悉的一些东西。

在实际操作中，这一过程可以这样进行：对每一单位时间，每个个体已知处于风险中，这样就可以建立一份独立的观测数据。在我们的生物化学家案例中，个体以人次记录而时间是以年记录，我们可以将这些观测视为人—年(person-years)。因此，在第 1 年升迁的

生物化学家对样本贡献了 1 人—年,而在第 3 年升迁的个体贡献了 3 人—年。被删截的个体——在 10 年中离开学术界或没有升迁的个体——贡献了最大的 10 人—年。因此对于 301 个生物化学家,总共有 1 741 人—年。从表 2.1 可以看出,这一总体恰恰是 10 年时间里处于风险中的数量之和。

对于每个人—年,如果个体在那一年升迁了,因变量记为 1,否则为 0。解释变量被赋予了他们在每人—年时的观测值。然而在有些情况下,研究者倾向于使用时变变量的滞后值。最后一步是将这 1 741 人—年放入一个单独的样本中,然后利用极大似然法估计一个二分因变量的 logistic 模型。目前主流的统计软件包都有进行 logistic 回归分析的统计程序。

要注意删截和时变解释变量这两个问题在这一过程中如何被解决。在这一处理过程中,没有升迁而被删截的个体准确告诉了我们应该知道的信息,换句话说,他们在被观测的任何年份里都没有升迁。时变的解释变量很容易被包括在内,因为处于风险的每一年都被单独观测。

第5节 | 生物化学例子的估计值

让我们来看将这一方法应用到生物化学数据时会有什么发现。表2.2报告了模型1的预测结果，这一模

表2.2 预测升迁概率的logistic模型，1741人—年

解释变量	模型1			模型2		
	b	*z*	*Exp*(*b*)	*b*	*z*	*Exp*(*b*)
本科院校（*undgrad*）	0.180	2.97**	1.20	0.194	3.05**	1.21
医学博士（*phdmed*）	−0.265	−1.64	0.77	−0.236	−1.37	0.79
博士声望（*phdprest*）	−0.003	−0.03	1.00	0.027	0.29	1.03
工作声望（*jobprest*）	−0.253	−2.40*	0.78	−0.253	−2.23*	0.78
文章数量（*arts*）	0.127	7.67**	1.13	0.073	4.05**	1.08
引用率（*cits*）	−0.001	−1.16	1.00	0.000	0.10	1.00
年 数（*Year*）				2.082	8.91**	8.02
年数平方				−0.159	−7.81**	0.85
截 距	−2.963			−8.484		
对数似然率	−595.57			−506.01		

注：* 在0.05的水平上显著，双尾检验。
** 在0.01的水平上显著，双尾检验。

型不允许机会随时间自发变化。(计算机命令可以从 http://www.statisticalhorizons.com/resources/books 中获取。)参数估计值像是非标准化回归系数,因为它们依赖于自变量的度量单位。我们首先进行 z 检验,零假设(null hypothesis)为每个系数都是 0。这些都是没有度量单位的并且给出了变量相对重要性的一些信息。

三个变量对升迁的机会有显著影响($p < 0.05$)。具体来说,毕业于高度选择性的本科院校和发表更多文章的生化学家有更高的升迁机会。另一方面,受雇于更知名的学术机构意味着更低的升迁机会。博士院系的声望、引用率以及从医学院毕业的影响并不显著。标记为 $Exp(b)$ 的一列给出的是指数化系数,这可以被视为几率比(odds ratio)。如果将几率比减 1,然后乘以 100,我们就能得到各个自变量每增加 1 个单位,升迁几率的百分比变化。对于 *undgrad*,每一分的增加(满分为 7 分),升迁的几率会增加 20%。受雇单位声望每增加一个单位,升迁的几率会降低 22%。最后,每多发表一篇文章意味着升迁的几率增加 13%。

模型 2 通过同时加入年份项和年份平方项从而允许机会随时间变化。两个变量的系数都是显著的。其他变量的结果并没有多少改变,但 *arts* 除外。它的 z 值明显下降(尽管仍然非常显著),并且几率比从 1.13 下降到 1.08。因此,在控制时间之后,每多发表一篇文章,升迁几率增加 8%。

第 6 节 | 似然比卡方检验

抛开其他变量，通过比较模型 1 和模型 2，我们可以检验受雇单位改变的机会不随时间变化这一零假设。这一过程非常类似于检验 R^2 增量的显著性（一般而言，这些增量来自被放到多元回归方程之中的额外变量）。只要某一模型是另一模型的特殊形式，这一检验即适用。例如，某一模型不仅包含了另一模型的所有变量，还加入了其他的变量。统计检验是极大似然估计的副产品，即对数似然函数的最大化取值。这在表 2.2 的两个模型中被分别给出。

为了比较两个模型的拟合度，我们可以计算它们的对数似然差异绝对值的 2 倍。（为了帮助这一统计量的计算，有些计算机程序报告的是对数似然的 −2 倍。）在无差异零假设的前提下，这一统计量呈现大样本卡方分布。相关自由度是比较两个模型的约束个数。在多数情形下，两个模型之间的差异仅仅体现在变量的数量上。在这个例子中，对数似然的差异的两倍是 179.12。因为模型 1 比模型 2 少了 2 个变量，也就

是 2 个自由度。这一卡方值远远高于 0.01 显著性水平上的取值。因此,机会随时间发生变化有很强的证据支持。

这一通过比较对数似然性以检验包含多变量集的假设所进行的步骤普遍应用于极大似然估计。因而,它同样适用于接下来章节要讨论的所有模型和估计过程。

除了检验时间的影响之外,研究者往往希望检验和时间的交互项。和普通线性回归和 logistic 回归类似,这可以通过在模型中加入交互项来实现。例如,我在模型 2 中加入 *phdprest* 和 *year* 的交互项,交互项的相关系数(−0.10)的 p 值是 0.04。*phdprest* 的主效应系数(0.57)的 p 值也是 0.04。这些结果显示博士学位的院系声望在学术工作的开始几年对升迁的几率有积极的影响,但是这一影响到第 6 年下降到 0。我们将在第 4 章看到,这类检验在方程上等价于对 Cox 回归模型比例机会假设是否被违背的检验。

第 7 节 | 离散时间的 logistic 方法存在的问题

离散时间方法的一个普遍问题是，由于个体在数据集中贡献了多个观察值，因此研究者需要找到一个能够矫正同一个个体的多个观测值之间相互依赖的方法。这类方法包括稳健标准误、一般化估计方程和随机效应（混合效应）模型。事实上，只要个人发生了不超过一次事件，那么研究者就没有必要进行修正（D'Agostino, Lee, Belanger, Cupples, Anderson & Kannel, 1990）。附录给出了简单的证明。完整的证明参见埃里森（Allison, 1982）。如果事件是可重复的，那么对相互依赖进行修正是必需的。

在生化学家的例子中，出于计算的考虑，构建人—年的数量有着很好的可操作性。但是，当一个在较长的时间跨度内追踪的大样本被分割为很小的离散时间单位时，构建观测值的数量将是难以计算的。在生物化学的例子中，将人—年转变为人—天（person-days），样本量将达到 635 000。研究者可以通过将数据整合

进更大的时间跨度来解决这一问题,但是这会不可避免地丢弃一些信息。(然而,值得注意的是,这个例子几乎不存在信息缺失,因为一般而言,大多数学术升迁发生在每学年的开始。)超越解释变量被测量的时间跨度来生成数据往往是不明智的。例如,如果解释变量是按月测量的,那么整合成人—年数据并不是好主意。

有几种途径可以将离散时间方法以最小的代价应用于大样本。例如,如果所有的解释变量都是定类的(或者有很少的离散值),基于因变量和所有解释变量的交叉分类,我们可以通过分组的数据来对 logistic 模型进行估计。在这一情形下,计算时间取决于列联表的格数而非所有的观测数。

另一个减少离散时间方法损失的途径是通过因变量来抽样。为离散时间的 logistic 回归所构建的数据集通常有相对较少发生的事件,但却有大量没有发生的事件。例如,在 1 741 人—年的生物化学样本中,仅有 217 个升迁事件,大概占 12%。我们的设计是:使用所有经历事件的样本,并且在没有经历事件的样本中进行简单的随即抽样,然后对这个重新组合的子样本进行 logistic 回归。这种基于因变量的不成比例的分层抽样并不会对 logistic 回归模型中的系数造成偏误(Prentice & Pyke, 1979)。显然,这里存在一些信息损失,导致了更大的标准误,但是这种损失通常是很小的。没有发生事件的样本越多,结果会越好,但是回报

会迅速降低。一旦每个经历事件的个案对应于5个未经历的个案，那么继续增加未经历事件的个案就意义不大了。

尽管方程2.2中的logistic模型是解决离散时间机会对解释变量依赖性最常用的方法，我们仍然有一种颇有吸引力的替代方法——互补的log—log模型：

$$\log[-\log(1-P(t))]=b_0+b_1x_1+b_2x_2(t) \qquad [2.4]$$

类似于logistic模型，这个模型确保了$P(t)$在0和1之间，不管方程的右边是什么。不同之处在于这个模型等价于连续时间数据的Cox比例机会模型，这会在第4章详细叙述。如果对方程2.4中的相关系数b进行指数化，那么得到的将是机会比而不是几率比。许多软件包提供了互补的log—log模型分析工具。在实际操作中，logistic模型和互补的log—log模型的选择并不是决定性的，尤其是如果研究者主要关注p值。互补的log—log方法的系数一般而言要比logistic回归得到的系数略小，但是这仅仅是因为模型量度上的差异，并不会有任何实际的差别。

第8节 | 删截

在这个例子中,删截发生的原因有两种:个体在升迁之前离开学术界,或者他们在10年之后仍然没有得到升迁。这两类删截都被称为右删截,也就是在个体被关注的最后一个时间点,事件仍然没有发生。但是这两种右删截之间存在着重要的差别。对于在10年之后仍然没有升迁的个体,这被称为固定删截,因为删截时间是由研究设计固定下来的,并且删截时间对于那些被删截的个体来说是一样的。

另一方面,对于离开学术界的人,正如表2.1所示,删截时间是变化的。如果删截时间对于不同的人是不一样的(并且不在研究者的控制之中),那么删截是随机的。当个体因为种种原因在观察期结束之前就退出了研究,这种情况便会发生。可能的原因包括死亡、人口迁徙、后续访谈中无法定位或者个体拒绝继续处于研究之中。随机删截也包括对所有个体的观察同时结束,但是不同的个体开始进入研究的时间不同。

不要被随机这个术语所迷惑。它并不意味着删截

和其他一切都没有关系。实际上,它的反面也可能是随机删截。我们说删截是随机的,其实是说它属于被研究现象的一部分,而不是研究设计的一部分。

当删截是随机的时候,事实上所有的事件史方法都假设删截时间并不能提供有效信息。这意味着在特定时间点被删截的个体并不能提供其在此时间点所面临机会的信息(在控制了模型中其他解释变量之后)。在我们的例子中,如果个人离开学术界是因为自身升职的机会很低,那么随机假设会被违背。基于我们所了解的晋升制度,其中一些人离开学术界就是因为自己无法升职,从而导致他们的学术职位被撤销。

遗憾的是,并没有方法能检验删截不能提供信息的假设(Tsiatis, 1975)。即便我们确信这一假设被违背了,但是没有有效的方法能够放松这一假设。所以大多数研究者往往忽略这一问题并抱乐观的希望。但是,研究者唯一能做的是他们的研究设计和执行能使随机删截的问题最小化。你不能简单地以常规的事件史方法来解决所有的删截问题。

尽管我们不能直接检验无信息假设,但是我们仍然能够判断一个研究在多大程度上违背这一假设。敏感性检验包括两次重新估计模型,每次都将按一种不同而极端的方法随机删截样本。在情形 A 中,随机删截的个案在被删截的时间点经历了一次事件,然后对这个数据重新进行模型估计。对于我们的离散时间方

法而言,这意味着将随机删截个案的最后一个时间点上的因变量从 0 变为 1。这一情形对应于随机删截的个案都是高机会的假设——这些个案机会很高以至于在删截时间点上经历了事件。

在情形 B 中,所有随机删截的个案的删截时间都被记录为数据中最大的观测时间点,不管是发生事件的时间还是删截的时间。因此,在我们升职的例子中,删截时间被设定为 10 年。(对于离散时间方法,这要求生成在删截时间点之外的额外记录,因变量被记为 0。并且,对于这些额外记录,任何时变解释变量的取值都是它们最后的观测值。)情形 B 对应的是随机删截的个案有很低的机会——低到即便他们经历了在风险中数据观测的最大时间,他们仍然没有经历事件。

如果标准分析中的参数估计(以及它们的统计显著性)和这两种极端情形下的结果相似,那么研究者可以认为违反无信息假设并没有严重的影响(Peterson, 1976)。注意,这一方法同样适用于接下来几章讨论的方法。同样值得注意的是,这种调整数据的方法仅仅适用于随机删截的个案,而非固定删截的个案。

表 2.3 给出了在两种极端情形下升职模型的变化。为了便于比较,标注为"标准分析"的面板显示的是表 2.2 中模型 2 的结果。*undgrad* 和 *arts* 的影响在三种情形下基本一致。但是 *jobprest* 的影响在情形 B 下大大减小并且在统计上不再显著。*phdmed* 的影响在情

形A下是最大的并且在统计上显著，但是在其他情形下并不显著。需要注意的是，情形B(删截个案的升职机会很低)要比情形A更加合理。尽管表2.3显示的是每种模型的对数似然率，但是我们并不能进行模型间的比较。为什么？因为只要你改变了数据，那么对数似然率便不能进行比较。

表2.3　信息性删截的极端情形

解释变量	标准分析		情形A		情形B	
	Exp(*b*)	*z*	*Exp*(*b*)	*z*	*Exp*(*b*)	*z*
本科院校(*undgrad*)	1.21	3.05**	1.15	2.43*	1.20	2.90**
医学博士(*phdmed*)	0.79	−1.37	0.70	−2.29*	0.91	−0.57
博士声望(*phdprest*)	1.03	0.29	0.92	−0.97	1.14	1.43
工作声望(*jobprest*)	0.78	−2.23*	0.68	−3.73**	0.88	−1.24
文章数量(*arts*)	1.08	4.05**	1.05	2.88*	1.10	5.75**
引用率(*cits*)	1.00	0.10	1.00	1.28	1.00	−1.14
年　数(*Year*)	8.02	8.91**	5.54	9.48**	8.55	9.31**
年数平方	0.85	−7.81**	0.88	−8.04**	0.84	−8.85**
对数似然	−506.01		−608.28		−553.13	

注：* 在0.05的水平上显著，双尾检验。
** 在0.01的水平上显著，双尾检验。

第 9 节 | 离散时间 vs.连续时间

在转向连续时间方法之前,必须要强调此处论述的离散时间所得到的结果看起来和后面所讨论的连续时间方法产生的结果非常相似。实际上,当时间单位越来越小时,方程 2.3 的离散时间模型将转换为第 4 章所提及的比例机会模型(D'Agostino et al., 1990)。尽管不知道事件发生的确切时间产生了一定的信息缺失,但这种缺失对估计的标准误无关紧要。

因此,离散时间方法和连续时间方法之间的选择通常应基于计算的成本和便利度。如果没有时变解释变量,那么采用下面两章所描述的方法去做事件史分析会更为简便。这主要是因为连续时间方法并不要求每个个体的观测时间划分为一组不同的观测时间单位。相反,如果存在时变解释变量,那么使用连续时间方法和离散时间方法的成本和便利度就值得比较了。

第3章

连续时间数据的参数法

尽管第 2 章中的离散时间方法有着广泛的应用，但是大部分事件史分析是借助连续时间方法完成的。在这一章中，我们将审视几种更流行的参数方法，它们主要应用于事件时间被精确观测到的数据。这些方法被称为参数法是因为模型的每个方面都是被指定的，除了那些必须被估计的固定参数值。和上一章一样，我们将注意力限定在个体经历不超过一次事件并且所有事件都相同的情形。

事实上有许多相近的分析这类数据的方法，初学者在如何选择上往往感到无所适从。我会尝试对这一选择提供指南。尽管对这些方法的深入理解需要极大似然法和微积分(包括简单的常微分方程)的知识，但一般的数学基础就能让你成为一个娴熟的使用者。

第 1 节 | 连续时间的机会

首先我们来定义机会函数。在前面一章，离散时间机会被定义为，给定个体在时间 t 上面临着风险，个体在时间 t 经历事件的概率。然而，这一定义在连续时间下并不奏效，因为对每一个 t 而言，精确到时间 t 上的事件发生概率是无穷小的。取而代之，给定个体在时间 t 上面临风险，我们将这一概率视为从时间 t 到时间 $t+s$ 的间隔内，个体经历事件的概率，并将这一概率记为 $P(t, t+s)$。当 $s=1$ 时，它和第 2 章所定义的离散时间机会是相等的。

接下来，我们用这个概率除以 s(间隔的长度)，并且令 s 足够小使这一比例达到极限。这一极限就是连续时间机会，记为 $h(t)$。机会的其他常见符号有 $\lambda(t)$ 和 $r(t)$。形式上，

$$h(t)=\lim_{s\to 0}\frac{P(t, t+s)}{s} \qquad [3.1]$$

尽管方程 3.1 有助于我们得到事件发生的瞬时概率，但这并非是真实的概率，因为它可能大于 1。实际

上,它没有上限。更准确的解释是,$h(t)$是事件发生时未被观测到的比率。特别地,如果$h(t)$不随时间变化,例如$h(t)=1.25$,那么1.25是在一个单位长度的时间间隔内发生事件的期望数量。换句话说,$1/h(t)$是一个事件发生的期望时长,在这个例子中是0.80个时间单位。这一定义机会的方式和风险的直观概念相一致。例如,如果两个人分别面临0.5和1.5的机会,那么第二个人经历事件的风险是第一个人的3倍。

在大多数应用当中,假设机会随着某一时间函数而变化更为合理,无论时间是指最后一次事件发生以后的时间或者是个体的年龄。例如,有足够的证据显示,至少在25岁以后,被逮捕的机会随着年龄的增长而降低。相反,退休的机会随着年龄的增加而增加。由一些因素导致的死亡机会随年龄呈U形:在出生后相对较高,在早年迅速下降,但在中年以后开始上升。对机会函数的唯一限制是它不能为负。

需要重点理解的是,机会函数的形状是区分不同连续时间数据的分析模型的关键特征。实际上,机会函数$h(t)$完全决定了事件发生时间的概率分布(或者重复事件发生的时间间隔)。这一章接下来的部分我们会看到如何选择一个机会函数的形状。

同样值得注意的是,机会被定义为个体的特征,而不是总体的特征。有时我们假设每个个体有相同的机会函数,但是原则上,每个人都有一个独特的机会函数。

第2节 | 参数比例机会模型

有两类参数回归模型，比例机会模型（proportional hazards models）和加速失效时间模型（accelerated failure time models）。"比例机会"这一术语常常和 Cox 回归（在第 4 章讨论）相联系。但是 Cox 比例机会模型要比我们在这里讨论的参数模型更一般化。

通过机会函数，以及它如何依赖于时间和解释变量，比例机会模型可以最容易被确定。我们要考虑三个模型——指数模型、Weibull 模型和 Gompertz 模型，它们之间的差异仅仅是时间进入方程的方式不同。简便起见，我们假设只有两个解释变量，x_1 和 x_2，它们都不随时间而改变。一个最显而易见的方式是把 $h(t)$ 构建为解释变量的一个线性函数。但是，尴尬的地方在于 $h(t)$ 不能小于 0，但是我们没法指定线性函数不小于 0。那么一种典型的处理方法是在它等于解释变量的线性函数前取 $h(t)$ 的自然对数。因此，一个最简单的模型可以表述为：

$$\log h(t) = b_0 + b_1 x_1 + b_2 x_2 \qquad [3.2]$$

其中,b_0, b_1 和 b_2 都是被估计的常量。在这个方程中,$h(t)$是解释变量的一个函数,但并不依赖于时间。不随时间变动的机会暗示事件的发生时间服从指数分布,因此,它又经常被称为指数回归模型。

然而,指定一个恒定的机会通常是不现实的。例如,如果事件是死亡,由于器官的老化,这一机会会随时间增加。另一方面,如果事件是受雇者变动,随着时间增加,个体对工作的投入会增加,那么这一机会往往是呈下降趋势的。我们可以通过允许机会的对数值随时间线性地上升或下降来放宽恒定机会这一假设,即,

$$\log h(t) = b_0 + b_1 x_1 + b_2 x_2 + ct \qquad [3.3]$$

其中,c 是一个可正可负的常数。由于这一模型显示出事件发生时间的 Gompertz 分布,我们将方程 3.3 称为 Gompertz 回归模型。

另外,我们可以构建一个模型,使机会的对数值随着时间的对数值线性地上升或下降:

$$\log h(t) = b_0 + b_1 x_1 + b_2 x_2 + c \log t \qquad [3.4]$$

其中,我们规定 c 大于-1。这个模型产生了事件发生时间的 Weibull 分布。因此,它经常被称为 Weibull 回归模型。

还有很多其他模型,它们的差异只是在于时间进入方程的方式,但是上述三个模型是最常见的。然而,所有比例机会模型的特征在于并没有解释变量 x 和时

间的交互项。因此在方程 3.4 中，x_1 在所有的时间点都是一样的。关于参数模型的其他信息，请参见劳利斯(Lawless，2002)或者埃里森(Allison，2010)。尽管时间在 Weibull 和 Gompertz 模型中充当了解释变量，它的作用是更为基础性的。尤其要指出，方程 3.3 和方程 3.4 之间存在的差异要求一种完全不同的估计过程，而远非简单地从时间到时间对数的转换。

要注意的是，无论是 Weibull 模型还是 Gompertz 模型，都没有考虑机会随时间呈 U 形或倒 U 形，机会可能随时间下降或上升(或者保持不变)，但不会出现方向的变动。这在一些应用中会带来不便。之后我们会考虑一些没有这一限定的模型。

同样要注意的是，上述没有一个模型包括了随机干扰项。但它们并非确定性模型，因为未被观测的因变量 $h(t)$ 和观测到的事件时间之间的关系存在着随机变异。仍然有些学者认为这些模型应当包括随机干扰项。这一问题会在本章结尾进行讨论。

第 3 节 | 极大似然估计

写模型很容易,但困难在于对它们进行估计,尤其是使用删截数据的时候。在 20 世纪 60 年代末,统计学家提出了针对指数模型的极大似然估计法(Zippin & Armitage, 1966; Glasser, 1967),很快这一方法就应用到了其他许多模型中。附录第 1 部分中详细地讨论了参数模型的极大似然估计法,但有必要在这里介绍一些基本的知识。

作为删截数据的估计方法,极大似然估计令人难以拒绝。它结合删截和非删截的观测以得出渐近无偏的估计,这些估计服从正态分布并且是有效的(即有最小样本方差)。遗憾的是,"渐近"意味着它们只是随着样本量的加大而不断接近的近似值。在小样本中,这些近似值可能并不如我们所想的那样理想。但是,由于缺少可替代的方法,极大似然估计被广泛地应用于大样本和小样本。

许多统计软件包可以进行一种或多种比例机会模型的极大似然估计。它们主要是 JMP, LIMDEP, Minitab, R, SAS, Stata 和 Statistica,但不包括 SPSS。

第 4 节｜一个实证案例

为了说明这些方法，我们将指数回归模型应用到在第 1 章已经简单介绍过的犯人累犯的数据集（Rossi, Berk & Lenihan, 1980）。样本包括 432 个从马里兰州立监狱释放出的男性犯人在被释放一年之内的追踪数据。这一研究实际上是一个随机田野实验，其中一半犯人得到资金援助，而另一半则成为控制组。在随后的一年里，每月受访者都会汇报他们上个月的经历。在年底，研究者会搜集地方法院关于拘捕和审判的数据。

兴趣事件是释放后被首次逮捕，而研究目的是探究拘捕机会在多大程度上依赖于下述解释变量：年龄、种族、教育年限、婚姻状态、过往盗窃被捕的次数、假释状态、资金援助和过往的工作经历。在被追踪的一年里，上述所有变量都是常量。在第 4 章，我们会检验一个包括就业状态的模型，并允许其在一年的观测中随时间而变动。

大部分估计参数事件史模型的软件要求数据的因变量按照两个部分输入：表示在观测时期内事件（在这

一案例中是被捕)是否发生的虚拟变量,和给出事件发生时间(如果发生)或删截时间的变量。在这个例子中,时间是指自释放以来的周数。因此,对于那些被逮捕的人而言,因变量的第二个部分是指从释放到逮捕之间的时间间隔;而对于那些自始至终都没被逮捕的人来说,周数是代表一年的 52 周,即直到他们被观测的最后一周。在 432 个个案中,114 个在接下来的一年中被再次逮捕,而其他 318 个被删截。指数回归模型的估计可以通过 Stata 和 SAS 获得(程序列表可参见 www.statisticalhorizons.com/resources/books),结果参见表 3.1 的第 1 面板。

系数估计一般是非标准回归系数。例如,在控制其他变量的情况下,释放时年龄的系数−0.056 意味着年龄每增加一年,机会的对数降低 0.056。更直观的解释可以通过求系数的反对数得到。即,如果 b 是系数,计算 $\exp(b)$,即 e^b(e 约为 2.718)。这些指数系数参见表 3.1,常常被称为机会比,因为它们给出的是解释变量在一个数值上的机会和比这个数值低一个单位的机会的比例。例如,资金援助(一个虚拟变量)的机会比是 0.693。这意味着那些获得资金援助 ($x=1$) 被逮捕的估计机会是没有获得资金援助 ($x=0$) 被逮捕的机会的 0.693 倍。

对于量化的变量,计算 $100(\exp(b)-1)$ 代表解释变量每增加一个单位,机会发生变化的百分比(控制

表 3.1 三个累犯模型的估计

解释变量	1			2			3		
	指数模型			Weibull 模型			Gamma 模型		
	b	z	$Exp(b)$	b	z	$Exp(b)$	b	z	$Exp(b)$
资金援助(D)[a]	−0.366	−1.92	0.693	−0.382	−2.00*	0.682	−0.272	1.94	1.313
释放时年龄	−0.056	−2.55**	0.946	−0.057	−2.60**	0.944	0.041	2.46**	1.041
黑人(D)	0.305	0.99	1.356	−0.315	1.02	1.371	−0.225	0.99	0.798
工作经历(D)	−0.147	−0.69	0.863	−0.150	−0.70	0.861	0.107	0.65	1.113
已婚(D)	−0.427	−1.12	0.652	−0.437	−1.14	0.646	0.312	1.13	1.366
假释(D)	−0.083	−0.42	0.921	−0.083	−0.42	0.921	0.059	0.42	1.061
过往被捕	0.086	3.03**	1.089	0.092	3.22**	1.097	−0.066	3.09**	0.936
常数项	−0.147			−5.60			0.107		
对数似然		−325.83			−319.38			−319.38	

注：[a](D)表示虚拟变量。
* 在 0.05 水平上显著。
** 在 0.01 水平上显著。

其他变量)。例如,如果之前被逮捕次数的机会比是1.09,这告诉我们之前被捕次数每增加1次,机会估计增加9%。这同样适用于虚拟变量。资金援助的机会比是0.693,这可以转化为获得资金援助的个体被再次逮捕的机会下降31%。或者我们通过1/0.693得到1.44,可以说没有获得资金援助的个体被再次逮捕的机会比那些获得资金援助的个体高44%。

z 值通常是系数和它们的标准误的比。在大样本的情形下,以系数是0作为零假设,它们接近于标准正态分布。如果 z 值大于2且做双尾检验,系数在0.05的水平上显著。这些比例的大小也可以作为变量相对重要性的大致指标。在这个例子中,我们看到7个变量中仅有2个变量的影响是显著的:释放时年龄和以往被逮捕次数。资金援助的影响仅仅在单尾检验下显著但在双尾检验下不显著。所有这些影响都符合我们的预期。因此,以往被逮捕的正向影响意味着,那些以前有过被捕经历的人在任何时间点都有更高的被捕风险。

如前面所述,指数回归模型告诉我们机会会因个体而不同(由于他们解释变量的差异),但是对于个体而言,机会并不随时间变化。然而这太过于局限。在表3.1的第2面板,我们看到用Weibull回归模型进行分析的结果,它允许机会随时间升高或降低。结果和其他的指数模型差异不大,除了资金援助的系数变大了一点并且在统计上略微显著。

表格并没有给出方程 3.4 中时间的对数的相关系数 c，它的估计值是 0.404。这说明被捕机会在一年内是不断上升的。更具体来说，由于机会和时间都是取对数，我们可以说时间每提高一个百分点，被捕的机会会提高 0.404%（并不是 40%而是低于半个百分点）。

Weibull 模型要比指数模型好吗？由于指数模型是 Weibull 模型的一个特例，我们可以通过似然比检验来回答这一问题，就像第 2 章中的方法一样。将两个模型对数似然比的差乘以 2（表 3.1 所示），我们得到 12.90。在零假设下，简单的模型（指数模型）被认为更加合适，这一统计量服从 1 个自由度（因为 Weibull 模型有一个额外的参数 c）的卡方分布。在这个例子中，我们得到 p 值是 0.000 3，所以我们拒绝指数模型而选择 Weibull 模型。需要注意的是，这等价于检验 $c=0$ 的零假设。

第 5 节 | 加速失效时间模型

另一类常见的参数事件史模型被称为加速失效时间模型(Kalbfleisch & Prentice, 2002)。如果 T 是事件发生时已经消逝的时间,这类模型可以写为:

$$\log T = b_0 + b_1 x_1 + b_2 x_2 + u \qquad [3.5]$$

其中 u 是独立于变量 x 的随机干扰项,并且其常数方差是 σ^2。这本质上是一个以 T 的对数为因变量的传统的线性回归模型。这个模型名字“加速失效时间”来自这样的事实,即解释变量 x 会促进(或延缓)事件发生的时间(“失败”)。

这一类方法之间的差异在于随机干扰项 u 有不同的分布。这些分布包括正态分布、对数 gamma 分布,logistic 分布和极值分布,它们又分别对应 T 的对数正态分布、gamma 分布、对数 logistic 分布和 Weibull 分布。特别的是,这些模型反映的是 T 的分布而不是 u 的分布。事实上,Weibull 模型(和它的特殊情况——指数模型)既属于比例机会模型,也属于加速失效时间模型。

加速失效时间模型可以被重新表达，以便将因变量转换为我们所关注的机会而非 T 的对数，但这些表达往往会很复杂。和 Weibull 和 Gompertz 模型相反，对数正态模型和对数 logistic 模型将机会处理为时间的非单调函数；也就是，机会首先会随时间上升，达到一个顶点，然后渐渐下降。作为最一般化的模型，gamma 模型允许机会函数呈现出这些形状，甚至更多。尤其是，gamma 模型可以有一个 U 形的机会函数，它常常可以作为一个很好的死亡风险随年龄而变化的模型。

如果 T 不存在删截，那么加速失效时间模型可以很容易地通过普通最小二乘法(ordinary least squares)回归对 T 的对数进行估计，这会产生大致无偏的系数值。然而，在删截存在的情况下，研究者必须要借助极大似然估计。附录阐述了似然函数是如何被构建的。劳利斯(Lawless, 2002)给出了一个方程是如何被最大化的解释。

表 3.1 的第 3 面板报告了 gamma 模型对再次被捕数据的估计。新的结果中最令人意外的一个特点是，系数的符号完全和指数模型、Weibull 模型中的结果相反。这是由于模型构建的不同导致的。方程 3.5 中的 gamma 模型预测了事件时间的对数，而方程 3.4 中的 Weibull 模型预测的是机会的对数。机会低意味着距离事件发生的时间较长，而机会高意味着很短的时间。

尽管资金资助的 p 值再次稍微大于 0.05,但是 z 值大致相同。

gamma 模型的指数系数可以被解释为时间比。例如,对于资金援助,数值 1.313 告诉我们那些接受资金援助的个体被捕的期望时间要比没有接受资金援助的人长 31%。相似地,以往被捕经历的时间比为 0.936,说明以往的每次被捕会导致被逮捕的期望时间下降 6%。

哪个模型更好,是 gamma 模型?还是 Weibull 模型?Weibull 模型是 gamma 模型的一个特例(包含了一个额外的参数),所以我们可以通过似然比检验再次进行比较。但是要注意的是,表 3.1 给出的结果显示,两个模型的对数似然值是相等的,所以似然比的卡方是 0。我们不能拒绝更简单的 Weibull 模型而去接受更为复杂的 gamma 模型。因此我们选择更为简单的那个。

由于 Weibull 模型既是比例机会模型,又是加速失效时间模型,所以它可以通过转换得到和方程 3.5 一样的系数。如果这一例子这么做的话,得到的系数和表 3.1 报告的 gamma 模型的系数几乎相等。考虑到两个模型的拟合度很好,这个结果并不令人意外。

第 6 节 | 评估模型拟合度

研究者如何在可替代的参数模型之中选择？就大多数统计方法而言，很难编出一套涉及模型选择过程的宝典。在做决定时我们需要考虑很多因素，包括数学简便性、理论契合度以及实证证据。

我们首先考虑实证证据，因为它是最容易被量化的。由于所有的模型都是通过极大似然法估计的，我们可以采用极大对数似然值作为我们评估模型拟合度的基本测量。对数似然值越接近 0，拟合度越好。但由于对数似然值常常都是负的，为了更容易进行评价，我们经常把它们乘以 −2 倍。表 3.2 给出了目前我们讨论过的所有模型的 −2 倍对数似然值。

表 3.2 累犯模型的拟合优度

模　型	−2×对数似然	*AIC*	*BIC*
gamma	638.753	658.753	699.437
对数正态	645.389	663.389	700.005
对数 logistic	638.797	656.797	693.413
Weibull	638.753	656.753	693.369
Gompertz	641.199	659.199	695.815
指　数	651.652	667.652	700.199

值得注意的是，不同数据的对数似然率是无法进行比较的。这一方法仅仅适用于同一数据集的不同模型。而且，−2 倍对数似然率随着样本量的增加而增加，所以大数据集会有大统计数值。从表 3.2 我们可以看出 gamma 模型和 Weibull 模型有最低的值，因此是拟合度最好的模型。然而，Weibull 模型和对数 logistic 模型的差异很小。

如前所述，如果一个模型是另一个模型的特例，那么它们的−2 倍对数似然的差异就是似然比卡方统计量。我们已经计算了比较指数模型和 Weibull 模型(有显著差异)的卡方以及 Weibull 模型和 gamma 模型(没有显著差异)的卡方。但是我们在此也可以进行另一种比较，就是 gamma 模型和对数正态模型。卡方值是 645.389 − 638.753 = 6.636。它有 1 个自由度，p 值是 0.01。因此 gamma 模型显著好于对数正态模型。

−2 倍对数似然作为拟合度检验的一个局限是有更多变量的模型往往有更好的拟合度。为了解决这一问题，AIC 和 BIC 统计法(表 3.2 所示)惩罚了有更多参数的模型。AIC(Akaike 信息判别法)的计算方法是：

$$AIC = -2\log L + 2k \qquad [3.6]$$

其中 $\log L$ 是对数似然值，k 是模型中参数的数量。BIC(贝叶斯信息判别法)的计算方法是：

$$BIC = -2\log L + k\log n \qquad [3.7]$$

其中 n 是样本量。由于 $\log n$ 常常比 2 要大，因此 BIC 要比 AIC 的惩罚更为严重。

因为对数正态模型、对数 logistic、Weibull 模型和 Gompertz 模型都有相同的参数量，那么不管采用哪种拟合检验方法，模型保持着同样的排序。但值得注意的是，不论是采用 AIC 还是 BIC，gamma 模型的拟合度都要比 Weibull 模型差，尽管它们的对数似然表面上看是相等的，这是由于 gamma 模型多了一个参数。

比较这六个模型，显然 Weibull 模型是胜者，但是对数 logistic 模型并没有差很多。这很令人惊讶，因为 Weibull 模型告诉我们，机会要么增加，要么降低，但是并不会发生方向的变化。而对数 logistic 模型告诉我们，机会首先会升高，在达到顶点的时候会下降。我们可以通过检验每个模型的机会函数来更好地理解这些差异。

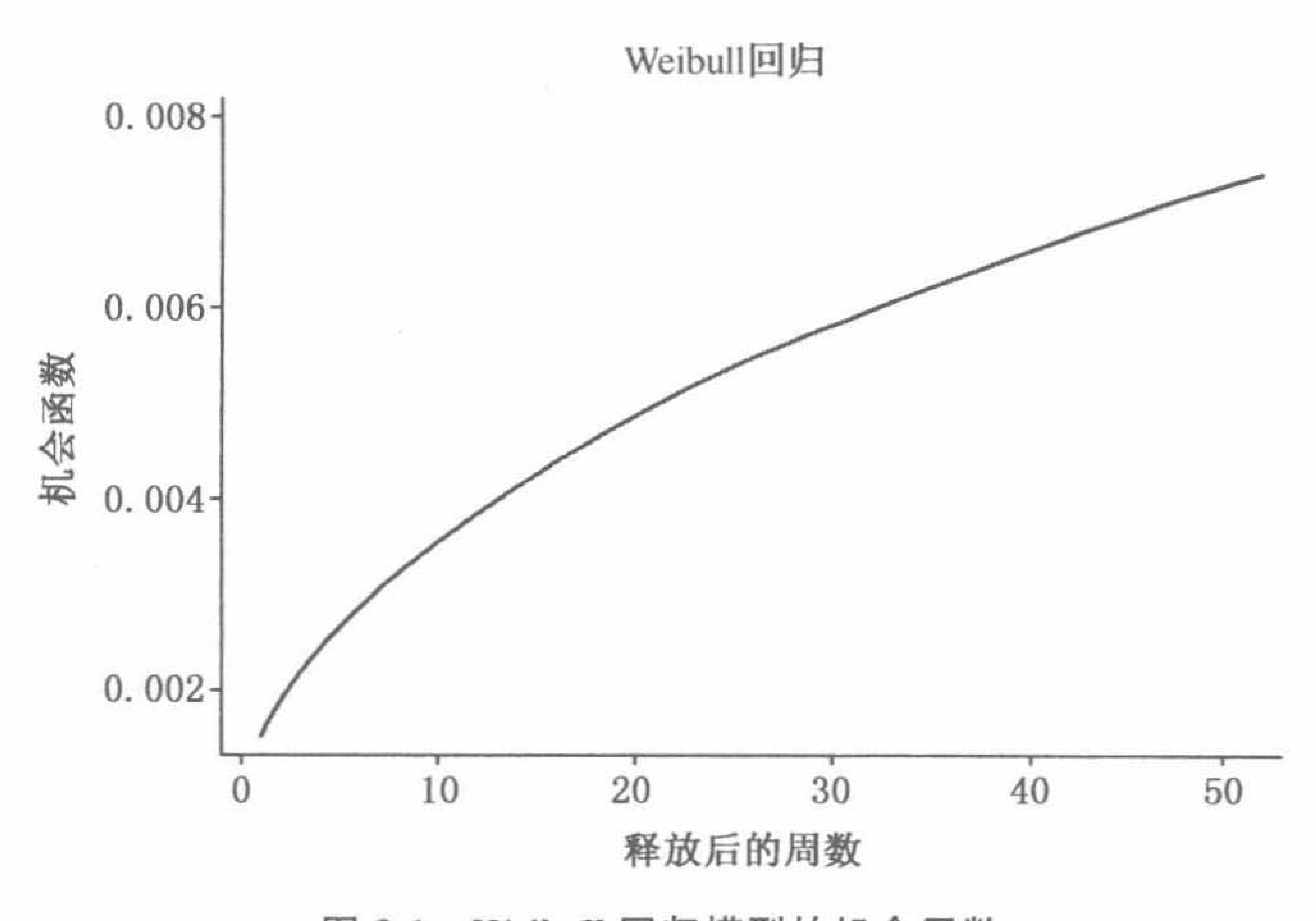


图 3.1　Weibull 回归模型的机会函数

图 3.1(在 Stata 中由 **stcurve** 命令生成)显示了 Weibull 模型的机会函数,所有的解释变量保持在均值。图 3.2 对应的是对数 logistic 模型的图形。在 52 周的研究中,两个曲线都在增长(增速递减)。理论上,对数 logistic 曲线最终会下降,但是我们手中的数据并不能验证这一可能性。

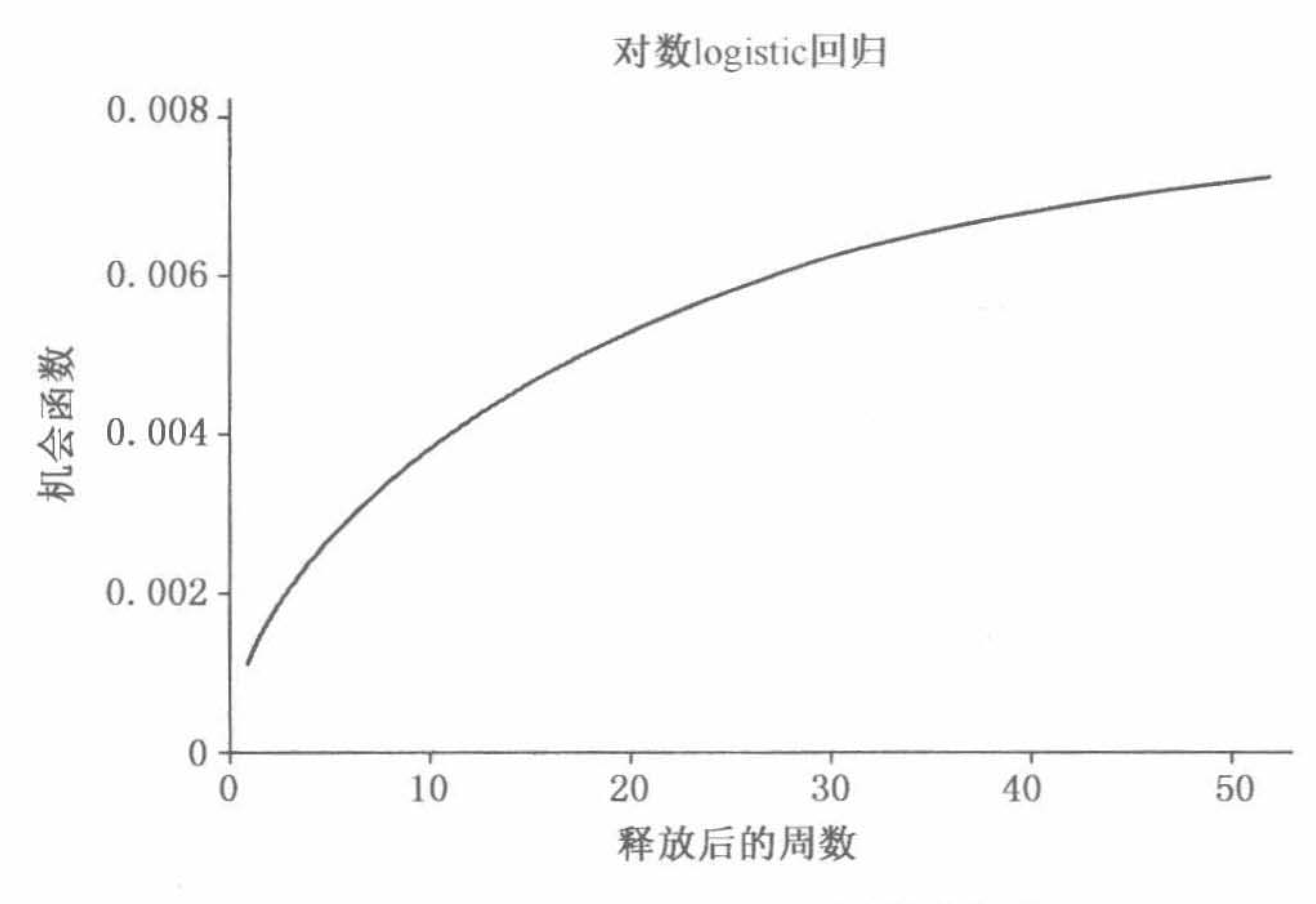


图 3.2　对数 logistic 回归模型的机会函数

第7节 | 异质性的隐性来源

很多理论暗示或认为某些事件发生的机会随时间增加或下降。例如，某些求职理论认为，获得工作的机会随失业时间的增加而增长。尽管在前面关于模型选择的章节里讨论的过程对检验这样的假设非常有用，但是我们仍要小心推断时间对机会的影响。基本的问题是：即使对每个个体而言，机会不随时间变动，但是解释变量中没有包含的个体间机会差异会产生机会不断下降的证据（Heckman & Singer，1982）。

直观地讲，机会高的个体往往更早地经历事件并且在随后被剔除出风险集。随着时间推移，这一选择过程会导致只包含那些先赋性地拥有较低风险的个体。结果是我们很难区分风险是真的随时间下降还是只是个体间机会差异带来的简单变异。另一方面，如果我们观测到一个不断上升的机会，那么一般认为机会确实是随着时间而增长。

处理这种异质性的最好方法是直接将异质性的来源作为解释变量纳入到模型之中。但是，假设所有异

质性来源都可以被测量和包括在内,这是不现实的。这一问题引导了一些学者(例如 Heckman & Singer, 1982)将参数机会模型扩展为包含一个代表异质性不可观测的来源的干扰项的模型。例如,考虑一个包含随机误差项的 Weibull 模型:

$$\log h(t) = b_0 + b_1x_1 + b_2x_2 + c\log t + e \quad [3.8]$$

其中 e 是随机干扰项。原则上,这一模型的估计允许研究者将时间效应和未被观测到的异质性区分开来。遗憾的是,我们经常发现 c 和 b 系数的估计对 e 分布的选择和依赖于时间的函数形式(例如 t 或 $\log t$)的选择高度敏感。尽管许多软件包允许我们分析这一模型(通常被称为异质模型),但是我并不建议使用这些软件包,除非遇到以下两种情景:(1)当个体存在重复事件时,这在第 6 章会有描述;或者(2)当个体从属于一个群体,这个群体中的所有成员的 e 值都相同的时候。在这些例子中,研究者可以估计具备较好统计特征的共享异质模型。

第 8 节 | 为什么选择参数模型？

在下一章，我们会检验 Cox 比例机会模型，这也是进行事件史回归分析最流行的方法。我们将会看到，Cox 模型有许多吸引人的特点。但是，研究者为什么会希望使用参数模型呢？参数模型有两个 Cox 模型并不具备的优点。第一，参数模型可以很好地处理左删截和内部删截。左删截是指个体已经经历了一次事件，但是我们并不知道它什么时候发生的。例如，如果我们知道一个女性在 20 岁之前结过婚，但是我们并不知道她结婚时的具体年龄，那么她的结婚年龄就被左删截了。内部删截是指我们知道一个事件发生在两个时间点之间（比如说 20 岁和 30 岁之间），但是我们并不知道它在这一区间发生的具体时间。Cox 回归无法处理左删截。而对于内部删截，它也仅仅只能处理一些特例，例如区间对于样本内的所有个体都是一样的。所以，如果你的数据存在这两种删截，那么参数模型是非常有吸引力的。

Cox 模型同样不擅长于生成预测值，我们将在下

一章看到;但是参数模型可以生成预测值。原则上,你可以得到任何你想预测的值:事件的期望时间、事件的发生时间中位值、其他分位值或者整个的概率分布。所以如果你想要的是预期事件发生的时间的话,参数模型是合理的选择。但是这存在一个危险,参数模型产生的预测很可能会超出实际的观测时间。这种超出值和模型有关,并且我们往往不能确信我们的模型是正确的。

直到最近,Cox 回归在处理时变解释变量上表现得更好(下章将会解释)。大多数处理参数模型的软件包不允许时变变量。但是,现在起码有两个软件包(LIMDEP 和 Stata)能够估计有时变变量的参数模型,而且它们和 Cox 回归一样简单易懂。



Cox 回归

第3章讨论的参数方法能非常有效地分析事件史数据,但它们仍然存在一些缺陷。首先,我们有必要决定这些模型哪一个更加适用,而这个决定往往很难并且不确定。其次,大多数用于参数回归模型的软件不允许解释变量随时间变化。

这些问题在1972年得到解决。大卫·考克斯(David Cox),一名英国统计学家,发表了一篇名为《回归分析和生命表》的论文。在文中,考克斯提出了直到现在依然非常流行的模型和估计方法。

第 1 节 | 比例机会模型

考克斯的模型通常被称为“比例机会模型”(proportional hazard model),是我们刚才讨论的那些参数比例机会模型的简单一般化形式。我们姑且搁置时变解释变量的模型,先来看两个不随时间变化的变量,这一模型可以写为:

$$\log h(t) = a(t) + b_1 x_1 + b_2 x_2 \qquad [4.1]$$

其中 $a(t)$是任何可能的时间函数。由于这个函数没有被特别指定,模型通常被描述为半参数的或部分参数的。之所以称其为比例机会模型是因为对于任何时间点的任何两个个体而言,他们的机会之比是常数。形式上,对于时间 t,$h_i(t)/h_j(t) = c$,其中 i 和 j 分别代表不同的个体,c 可能取决于解释变量而并不依赖于时间。这一名称并非是 Cox 模型的关键特征,因为我们将会看到,这一模型可以很容易地被拓展为针对非比例的机会。

很容易发现第 3 章的参数比例机会模型是这个模型的特例。如果 $a(t)$是常数,我们会得到指数模型。

Gompertz模型将 $a(t)$ 定义为 ct，而 Weibull 模型有 $a(t)=c\log t$。但是还存在很多其他的可能性，例如四阶多项式：

$$a(t)=a_0+a_1t+a_2t^2+a_3t^3+a_4t^4 \qquad [4.2]$$

Cox模型优点在于 $a(t)$ 可以是任意形式，我们没必要给予特别的关注。

第 2 节 | 部分似然

写模型容易，但是找到好的方法去估计它们却不易。考克斯最重要的贡献是提出了一个名为部分似然的模型，这一模型在很多方面和普通极大似然估计相似。关于部分似然的数学表述参见附录第 1 部分，这里仅提及它的基本特性。这一方法依赖于以下事实，即处理基于比例机会模型的数据的似然函数可以被分为两个要素：要素一包含仅仅涉及系数 b_1 和 b_2 的信息；要素二包括关于系数 b_1 和 b_2 的信息以及函数 $a(t)$。部分似然简单地抛弃了第二个要素并且将第一个要素视为普通似然函数。标准的数学方法可以通过最大化部分似然来找到 b_1 和 b_2 的值。

部分似然一个有意思的特点是它仅仅依赖于事件发生的顺序，而并非事件发生的准确时间。在做 Cox 回归分析的时候，这一特点可以简单地理解为对时间变量的转化。在保留事件发生时间顺序和删截时间的前提下，任何转化都会带来相同的估计值和标准误。例如，你可以取时间的平方、时间的对数以及将时间乘

以一个常数,结果都不会改变。

部分似然估计值和极大似然估计值的三个特点中有两个相关,即一致性(因此在大样本里接近无偏)和正态分布(通过重复样本)。但是它们并不十分有效,这是因为它们有着比极大似然估计值更大的抽样变异(更大的真实标准误)。这是由于在忽略事件发生的具体时间时,一些信息被丢弃了。然而,这种效率的降低往往很小,所以不足为虑(Efron, 1977)。

Cox 回归现在一般被定义为比例机会模型的部分似然估计。考克斯的工作在事件史分析领域的影响十分巨大。近年来,他在 1972 年发表的那篇文章每年被来自全世界的科学研究引用超过 1 000 次。这一数字其实远远低估了 Cox 回归的运用,因为大部分研究者都不再引用最初的那篇论文了。在众多的学者眼中,Cox 回归无疑是用回归模型估计事件史数据最好的方法。目前所有的主流软件包都支持 Cox 回归的分析。

第3节 | 部分似然应用于累犯数据

我们现在将Cox回归方法应用于我们第3章分析的累犯数据。使用与指数模型和Weibull回归模型相同的解释变量，部分似然估计可以通过Stata中的**stcox**命令和SAS中的PHREG过程来分析。结果在表4.1的第1面板。

在表4.1中，估计的系数和z值与表3.1中的Weibull回归模型的结果基本相等。这并不奇怪，因为它们都是比例机会模型，但有时你会看到更大的差异。指数化的系数可以理解为机会比。例如，对于资金援助，0.684的风险比告诉我们获得资金援助的人被再次逮捕的机会比那些没有获得资金援助的人要低32%。相似地，对于以往的被捕经历，其每增加一次，被捕风险会提高9.5%。

需要注意的是，我们没有看到截距项。这是Cox回归的特点，因为截距已经是未被指定的函数$a(t)$的一部分，其在部分似然中完全消失。截距项的缺失也是Cox模型不便于进行预测的一个原因。但是我们稍后会在本章中找到一种解决办法。

表 4.1　累犯数据的 Cox 回归估计

解释变量	1			2			3		
	基本模型			包含时变变量 X			滞后 X		
	b	z	$Exp(b)$	b	z	$Exp(b)$	b	z	$Exp(b)$
资金援助(D)[a]	−0.379	−1.98*	0.684	−0.356	−1.86	0.700	−0.351	−1.83	0.704
释放时年龄	−0.057	−2.60**	0.944	−0.046	−2.12*	0.955	0.050	−2.27*	0.952
黑人(D)	0.314	1.02	1.369	0.339	1.09	1.403	0.322	1.04	1.379
工作经历(D)	−0.151	−0.71	0.860	−0.027	−0.13	0.973	−0.049 2	−0.23	0.952
已婚(D)	−0.433	−1.13	0.649	−0.293	−0.76	0.746	−0.344 2	−0.90	0.709
假释(D)	−0.085	−0.43	0.918	−0.064	−0.33	0.938	−0.047	−0.24	0.954
过往被捕	0.091	3.18**	1.095	0.085	2.93**	1.088	0.092	3.18**	1.096
就业(D)				−1.32	−5.28**	0.266	−0.782	−3.59**	0.457
对数似然		−659.12			−641.54			−646.17	

注：[a](D)表示虚拟变量。
* 在 0.05 水平上显著。
** 在 0.01 水平上显著。

第4节 | 时变解释变量

Cox比例机会模型可以很容易地扩展为允许时变解释变量的模型。一个包含两个解释变量(一个时变，一个非时变)的模型可以写为：

$$\log h(t) = a(t) + b_1 x_1 + b_2 x_2(t) \qquad [4.3]$$

其中，如前所述，$a(t)$可以是任何形式的时间函数。这一模型告诉我们，在时间t上的机会依赖于同一时间t上的x_2的取值。然而，在一些情形下，可能有理由认为变量的变化在其对机会的影响上存在延迟。例如，如果研究一次失业对离婚机会的影响，有理由怀疑失业这一事件的发生和离婚机会增加之间存在时间间隔。如果间隔被推测为1个月(时间单位是月)，那么模型应该修订为：

$$\log h(t) = a(t) + b_1 x_1 + b_2 x_2(t-1) \qquad [4.4]$$

无论有无时间滞后，包含时变解释变量(也被称为时变协变量)的模型可以通过前面所论述的部分似然方法进行估计。部分似然函数的推导过程本质上和包

含时变解释变量时是一样的,但是在计算机运算中,构建和最大化这一似然函数的计算过程却更为复杂。此外,并不是所有的部分似然函数估计的程序都支持时变解释变量。

让我们看看如果把一个时变解释变量加入到累犯的例子当中。如第 1 章所述,数据包含了 52 个虚拟变量,这些虚拟变量表示个体是否在 52 周的每一周内被雇佣。我们估计方程 4.3 的模型,这一模型认为个体在第 t 周被捕的机会依赖于同一周内个体的工作状态。在下一部分我们会考虑实现这一分析的技术细节,但是现在我们先来看表 4.1 中第 2 面板的结果。

对于已经包含到模型里的变量,第 1 面板和第 2 面板的结果很类似。最显著的不同在于资金援助的 p 值在第 1 面板稍稍低于 0.05,但是在第 2 面板稍稍大于 0.05。真正的大变化是,工作状态现在成为了模型中最重要的变量,不仅是因为它的影响的大小,而且也是由于其显著度。0.266 的机会比告诉我们,在特定周,有工作的个体被捕的机会要比没有工作的个体的机会低 73%。也就是说,失业的个体被捕的机会比就业的个体高 4 倍。

但是有一个问题,就业对被捕机会的影响可能是反向因果关系。如果一个人在一周开始的时候被捕,那么这就会对其在那一周剩下的时间里是否工作产生巨大的影响。为了避免这一问题,我们将就业状态延

迟一周。也就是说第 15 周被捕不会影响第 14 周的就业状态。表 4.1 的第 3 面板显示使用延迟一周的工作状态的结果。就业影响的大小显著下降，但其仍然保持了所有变量中最大的 z 值。根据这个模型，在给定一周工作的情况下，(下一周内)被捕机会下降一半。就业“真实”的因果效应可能是在第 2 面板和第 3 面板的结果之间。但是，在第 3 面板使用滞后变量的情况下，我们排除了其他可能的解释。

第 5 节 | 应用包含时变解释变量的模型

既然我们已经看到了包含时变解释变量的 Cox 模型的一个例子,那么我们可以关注一下更为技术化的细节。这类模型的数据准备和软件编程往往很复杂,研究者需要保持谨慎,因为错误经常发生。

时变解释变量的一个重要问题是这些变量被测量的频次。严格意义上讲,这些模型的估计要求每次事件发生时,解释变量的值对于所有处于风险中的个体都是已知的。因此,如果一个事件发生在时间点 10 且有 15 个个体处于风险中,那么在时间点 10 上的解释变量的取值对全部 15 个个体而言必须是已知的。遗憾的是,由于我们往往不能预先知道事件发生的时间或者谁会处于风险集之中,这就要求我们必须知道在每一时间点上每个个体的解释变量的值。

然而,在实际中,时变解释变量往往是在固定的时间间隔上被测量的。在刚刚提到的例子中,工作状态

的信息是以每周的观察被获得的。这并没有问题，因为被捕事件也是以周为时间单位被观测的。但是，当事件时间的测量比解释变量的测量在时间上更为精确的时候，问题就产生了。例如，假设事件发生时间是按天测量的，而解释变量的取值只是在每个月初获得。例如，如果一个事件发生在 5 月 17 日，我们需要知道每个处于风险中的人在那一天的解释变量值，但事实上我们只有 5 月 1 日和 6 月 1 日的值。

在这些例子中，有一些特别的步骤可以用来“计算”事件发生时解释变量的值。一种方法是使用和事件发生时间最接近的值。另一种方法是，研究者可以使用事件发生之前和发生之后通过线性插补测量的值。但是，这些方法都存在逆向因果的问题，因为其使用的信息都是事件发生之后搜集的。如果考虑这个问题，最安全的方法是计算最接近事件发生时间之前的取值。这一方法也被称作“最后值接续法”（last value carried forward）。

在一些必要的计算执行之后，下一步是决定如何构建数据以便于分析。对于包含时变变量的模型，实际上有两种截然不同的方法去构建数据和指定模型：分割区间法和编程陈述法。不同的软件包会使用这两种方法的某一种。例如 Stata 和 R 使用分割区间法，而 SPSS 只使用编程陈述法。SAS 提供了两种方法的选择。

编程陈述法

当研究者使用这一方法的时候,数据必须是"横向"(wide)的。也就是说,对每个人而言,只存在唯一的记录。时变变量必须编码为这一记录的多个变量。例如,在累犯数据的横向形式下,总共有 432 条记录,每个个体对应 1 条,52 个虚拟变量包含了 52 周观察的每一周内的就业状态的值。

在实际的 Cox 回归模型中,仅仅有一个变量代表每个时变解释变量。因此,在指定模型之后,研究者需要制定一个程序或算法把数据中的横向形式的变量转化为适用于模型的单个变量。这个步骤必须在进行部分似然分析的时候进行——不能在运行 Cox 模型之前进行。遗憾的是,我们很难找到一个普遍适用的指南来解释这一步骤,因为不同的软件包有不同的编程方法和独特的命令。这里给出了表 4.1 中第 2 面板的 SAS 命令:

```
proc phreg data = recid; model week * arrest(0)
    = fin age race wexp mar paro prio work;
array wrk( * ) w1-w52;
work = wrk[week];
```

在 SAS 中,进行 Cox 回归分析的步骤(proc)是通

过 phreg 这一命令完成的，它代表“比例机会回归”。在 model 陈述上，work 是不存在于原始数据集的新变量。SAS 期望发现新的程序陈述来定义这一新变量。接下来的两个陈述就是完成这一目的。array 的陈述定义了一个名为 wrk 的排列以包含 52 个虚拟变量：w1-w52。排列是指变量的有序集，它允许我们将特定的变量转化为一个排列。在这个排列里，名字后面的数字“下标”表示这些变量的顺序。下一个陈述将特定周 week 的虚拟变量的值赋予 work 这个变量。例如，如果 week＝10，那么变量 w10 的值就会赋予 work 这个变量。如果 week ＝ 38，那么 w38 的取值就会赋予 work。需要注意的是，在这一陈述中，week 这个变量并不是每个人被捕或删截的那个周，而是在构建部分似然函数时所依赖的变量。因此，如果一个人在第 38 个星期被捕，那么每个在第 38 周处于风险的人都会被赋予在部分似然估计中 w38 的取值。

分割区间法

这个也被称为计数过程法，这一方法要求数据被构建成“纵向”(long)结构，即每个人有多个记录。具体来说，对于每一个数据记录，必须对应于每个时间区间，并且所有解释变量都一致。因此，只要解释变量的值发生变动，当时的记录就会终止，新的记录就会产

生。从个体被观测的时间开始,每条记录都需要有起始时间和结束时间。如果记录并不是以事件发生而终止的,那么这条记录就被视为删截;如果它最终以事件发生而终止,那么该记录就是未删截的。

一旦数据成为纵向结构,我们可以将时变变量作为不随时间变化的变量——因为每一条记录上变量的值都是固定的。模型指定和我们前面所做的不同之处在于:对于每条记录,除了将变量的命名包含终止时间之外,我们还必须指定这个变量包含事件的起始时间。当然,不同的软件包在进行这个步骤时有不同的方法。

还有重要的一点。在解释变量取值发生变动的时候,我们必须将它分割为不同的记录,但是即便没有发生变动,我们也可以将它们分割为更小的时间区间。无论你将数据分割得多零碎,结果并不会有变化。这一点很重要,因为我们常常将数据分割为尽可能小的时间单位,而不是写一个命令去识别变量取值是否发生变化。例如,对于累犯数据,我将每个人从被释放到第一次被捕(或删截)的时间分为数个人—周。对于432 人的样本,我生成了一个包含 19 809 人—周的数据集。对于每个记录,start 表示开始时间,stop 表示结束时间。

表 4.2 呈现了数据的结构,它显示了数据中被捕的两个人的 13 条记录。表中没有给出不随时间变动的

变量。个体 339 在第 9 周被捕，因此其在数据集里贡献了 9 个记录：他在第 1 周到第 6 周并没有全职工作，只在第 7 周到第 9 周有全职工作。在被释放时，他 26 岁，并且获得了资金援助(fin＝1)。个体 417 的年龄是 18 岁并在第 4 周被捕，他在 4 周的时间里没有工作，同时也没有获得资金援助。

表 4.2　分割区间法的部分数据

编号	stop	start	arrest	work	fin	age
339	1	0	0	0	1	26
339	2	1	0	0	1	26
339	3	2	0	0	1	26
339	4	3	0	0	1	26
339	5	4	0	0	1	26
339	6	5	0	0	1	26
339	7	6	0	1	1	26
339	8	7	0	1	1	26
339	9	8	1	1	1	26
417	1	0	0	0	0	18
417	2	1	0	0	0	18
417	3	2	0	0	0	18
417	4	3	1	0	0	18

这里给出了生成表 4.1 第 2 面板的 Stata 命令：

```
stset stop, failure(arrest==1) timeø(start)
stcox fin age race wexp mar paro prio educ work
```

stset 命令定义 stop 为结束时间，start 为开始时间，arrest 为删截指示变量，1 表示事件发生。**stcox** 命令告诉 stata 估计一个包含解释变量的 Cox 模型。

两种方法的利弊

如果分析过程都是正确的话,那么两种方法的结果理论上应该是一样的。如果你忠实于某一个统计软件,那么你别无选择(除非你用的是 SAS)。权衡而言,我会选择分割区间方法,有以下理由:

- 它更为直观;
- 当你生成数据集的时候,你只需要做一次编程。而使用编程陈述法,程序需要作为每次 Cox 回归运行的一部分;
- 它更容易辨别和修改错误。使用程序陈述方法,你不会直观地看到你的程序的结果。而应用分割区间方法,你可以仔细观察“纵向”数据集以确保一切能正常运行。
- 方法本身会进行劳动分工。有经验的程序员会生成便于数据分析员进行分析的“纵向”数据集;
- 对于分割区间法,有很多诊断统计量。但是这些统计量并不适用于程序陈述法。

分割区间法的劣势在于,生成纵向数据集的命令有时会比程序陈述法更为复杂。但是,如前所述,你只需要做一次编程。

分割区间法所需要的数据集和第 2 章讨论的离散

时间方法所需要的数据有很大的相似之处。实际上，用两种方法分析同样的数据集也是可以的。我在第2章中对离散时间方法的讲述同样适用于此。尽管数据集里的每个个体都有多条记录，我们没有必要执行额外的步骤去修正观察值之间的相互依赖。也就是说，我们没必要去计算稳健标准误、随机效应、GEE或者固定效应。

第6节 | 检验和放松比例机会假设

许多研究者担心他们的数据是否满足比例机会假设。对那些有这种担心的学者来说,有很多方法用于评估这一假设是否被违背并且调整模型以适应非比例机会。但是,在讨论这些方法之前,我们首先考虑这些担心是否被夸大了。与前述章节的参数模型相比,比例机会模型是非常一般化和非限制性的,它也因此非常普及。即便没有满足比例机会假设,也会极为接近满足假设。那些担心可能会错误设定模型的学者其实更应该把注意力集中到可能被忽略的解释变量、解释变量的测量误差以及信息的删截上。

将上面所言谨记在心,让我们来讨论非比例机会的情况。机会是非比例的意味着什么?比例机会假设最重要的启示是,每个变量(对风险对数)的影响在所有时间点上都是一样的。因此,机会是非比例的主要原因是时间和一个或多个解释变量之间存在交互,例如

$$\log h(t) = a(t) + bx + cxt \qquad [4.5]$$

这个模型和其它模型的不同之处在于它将 x 和 t 的乘积作为一个解释变量。如果系数 c 是正数，我们说时间对机会的影响随着 x 的增加而线性地增加。因此，如果一些变量在不同的时间点对机会的影响是不同的，那么机会就是非比例的。

有很多方法可以检验这一假设。最简单的方法是基于 Schoenfeld 残差，很多 Cox 回归包里都有这一命令。这些残差（它们有比较抽象的定义）对每个经历事件的记录和每个解释变量进行计算（Schoenfeld, 1982）。如果机会确实是成比例的，那么 Schoenfeld 残差应该和时间或者时间的任意函数无关。

我们用这个方法来分析表 4.1 的第 2 面板。Stata 的命令 **estat phtest**, detail 报告了表 4.3 的结果。第一列包含了 Schoenfeld 残差和 114 个被捕周数的 Pearson 相关系数。对于每个相关系数，我们得到了对零假设（相关系数为 0）的卡方检验和 p 值。对于大多数变量来说，p 值都很高，这说明比例机会并没有问题。但是其中有两个变量，年龄和工作经历，它们的 p 值在 0.05 以下。这说明它们有可能违背了比例机会假设。也就是说，在一年的时间内，它们在不同的时间有不同的影响。在表格的底部，有一个对所有变量相关系数为 0 的全面检验，这个例子是在 0.03 的水平上显著。但是这个检验并不是主要的兴趣点，因为我们真正想知道的是每一个变量是否满足假设。

因为比例机会假设暗含着 Schoenfeld 残差和时间的所有函数形式无关,因此我们可以尝试除了时间这个变量本身之外的其他函数形式。Stata 提供了计算时间的对数、时间的顺序和事件发生时间的估计累积分布函数的相关系数的选择。我尝试了所有的函数,得到的结果和表 4.3 的类似。

表 4.3 比例机会假设的 Schoenfeld 残差检验

	相关系数	卡方	自由度	p 值
资金援助	0.04	0.21	1	0.648
被释放年龄	−0.27	11.68	1	0.001
黑　人	−0.11	1.42	1	0.233
工作经历	0.21	6.06	1	0.014
已　婚	0.06	0.49	1	0.483
假　释	−0.04	0.15	1	0.698
以往被指控	−0.00	0.00	1	0.988
就　业	0.04	0.18	1	0.668
总体检验		17.13	8	0.029

下一步呢?我们可以选择忽略潜在的问题,表 4.1 中第 2 面板里面的系数仅仅代表年龄和工作经历在一年时间里影响的大致平均值。这对于年龄来说已经足够好了(有一个很显著的系数−0.046),但对于工作经历来说,这可能不够好。尽管 Schoenfeld 检验显示存在和时间的交互作用,但是在 Cox 模型里工作经历的影响并不显著。这意味着工作经历在这一年里的一部分时间是积极的影响,而在其他时间里是消极的影响,两个影响相互抵消了。

对于比例机会假设可能存在的违背，一个解决方案是像方程4.5那样，把有问题的变量和时间的交互项纳入到模型之中。这个解释变量和时间的产物本身就是一个时变解释变量，那么我们可以使用我们之前讨论的方法进行分析，来把这些变量加入到模型中。如果统计软件使用的是程序陈述方法，这一分析就可以直接进行，但是对于分割区间方法有一点麻烦。举一个程序陈述法的例子，下面是SAS程序如何加入时间和年龄的交互项以及时间和工作经历的交互项：

```
proc phreg data = recid;
model week * arrest(0) = fin age race wexp mar
    paro prio work ageweek wexpweek;
array wrk( * ) w1-w52;
work = wrk[week];
agetime = age * week;
wexptime = wexp * week;
```

对于分割区间法，潜在的问题是时间是连续变动的，但是方法要求每个记录对应的是解释变量不发生变化的一段时间。对于一些软件包，唯一的解决方法是把区间分为相对较小的单位，每个区隔大小相等，就像我们处理累犯数据一样。然后，将时间看作每个区间里的常数，我们创建一个结合起始时间和解释变量的新变量。Stata有一个更好的方法来解决这个问题，

这一方法和程序陈述法很像。即便没有时变解释变量并且每个人仅仅有一个记录,这一方法仍然可行。**stcox** 命令有一个选项 **tvc**(时变协变量),它允许研究者构建和时间的交互项并把它们放进模型。下面是累犯数据的例子:

```
stset stop, failure(arrest==1) timeø(start)
stcox fin age race wexp mar paro prio work, tvc
    (age wesxp)
```

对于 SAS 和 Stata 来说,结果是一样的,如表 4.4 所示。

表 4.4 有时间交互项和分层的 Cox 回归估计

解释变量	有时间交互项			分　层		
	b	*z*	*Exp*(*b*)	*b*	*z*	*Exp*(*b*)
资金援助	−0.361	−1.89	0.679	−0.358	−1.87	0.699
释放时年龄	0.085	2.09**	1.089	−0.047	−2.17	0.954
黑　人	0.305	0.99	1.357	0.342	1.10	1.407
工作经历	−1.206	−2.50*	0.299	—	—	—
已　婚	−0.233	−0.61	0.792	−0.301	−0.79	0.740
假　释	−0.064	−0.33	0.938	−0.063	−0.32	0.939
过往被指控	0.080	2.78**	1.084	0.084	2.88	1.087
就　业	−1.314	−5.23**	0.269	−1.324	−5.28	0.266
年龄×周	−0.005	−3.29**	0.995			
工作经历×周	0.041	2.83**	1.042			

注:* 在 0.05 水平上显著。
** 在 0.01 水平上显著。

表 4.4 的最后两行,我们看到两个和时间的交互项都高度显著,这确认了我们从 Schoenfeld 残差检验中得

到的信息。这些交互项检验可以被认为比 Schoenfeld 残差更加明确。不过这些检验更美妙的地方在于诊断方法也能提供答案。我们现在有了一个允许年龄和工作经历是非比例机会的 Cox 模型。

尽管在交互项的解释上,这里和一般线性模型没有区别,但是比较有挑战性的一点是指出这些交互项到底意味着什么。解释应基于方程 4.5,这一方程可以写为以下形式:

$$\log h(t) = a(t) + (b + ct)x \qquad [4.6]$$

这一方程说明 x 的“影响”是时间的线性函数,其中 b 代表当 t 为 0 的时候 x 的效应,c 代表时间每增加一个单位,影响发生的改变。

应用到我们的累犯例子,年龄(0.085)和工作经历(−1.206)的“主效应”表示个体从监狱被释放的时候两个变量的影响。这些系数在统计上都显著。从对应的机会比(指数化系数)来看,我们了解到在被释放时候的年龄每增加 1 岁,被捕机会提高 9%,而有工作经历的人比没有这样经历的人被捕的机会要低 70%。

交互项系数反应的是这些变量在一年的时间里每增加一个周所发生的变动。对于年龄,每周对应的是下降 0.005 的被捕机会,而对于工作经历,则是上升 0.041 的机会。它们看起来都是比较小的变动,但是它们累加起来很惊人,如表 4.5 所示。年龄的影响开始是

正的,但是在第 20 周以后变为负。到第 50 周,年龄每增加 1 岁,对应的是被捕机会降低 15%。而对于工作经历,在这一年的前半个时期,它的影响是负的,在后半个时期,这一影响变为正的。到了第 50 周,那些有工作经历的人被再次逮捕的机会是那些没有工作经历的人的两倍。

表 4.5 不同时间年龄和工作的影响

周	年龄		工作经历	
	b	$Exp(b)$	b	$Exp(b)$
0	0.085	1.089	−1.206	0.299
10	0.035	1.036	−0.796	0.451
20	−0.015	0.985	−0.386	0.680
30	−0.065	0.937	0.024	1.024
40	−0.115	0.891	0.434	1.543
50	−0.165	0.847	0.844	2.326

非比例机会的交互项方法存在一个劣势,就是它要求交互项有特定的参数设定。我们假设时间对其他变量的系数有线性的影响,但是这没有什么可担心的。与此相反,例如,我们可以把时间的对数和工作经历的交互项纳入到模型之中。或者我们也可以将时间、时间的平方和工作经历的两个交互项纳入到模型之中。甚至我们可以将时间转化为二分变量,然后让一个解释变量在特定的时间之前和之后有不同的影响。

还有一种称为*分层*的方法可以让我们不做任何假设以允许非比例性。当不满足比例假设的变量是分类

变量的时候，这一方法尤其有用，比如说工作经历（二分变量）。下面是这一方法的应用。我们指定两个模型，分别对应于有工作经历的人和没有工作经历的人：

$$\text{有工作经历：} \log h(t) = a_1(t) + b_1 x_1 + b_2 x_2 + \cdots$$
$$\text{无工作经历：} \log h(t) = a_2(t) + b_1 x_1 + b_2 x_2 + \cdots \qquad [4.7]$$

这组方程有相同的相关系数 b，却有着不同的（但未指定的）时间函数。在时间 t 时，工作经历的影响是 $a_1(t) - a_2(t)$。在没有限定的前提下，在任何不同的时间 t 上，这一影响可以是不一样的。

系数 b 的部分似然可以很容易被构建以及最大化。软件首先分别构建对于有工作经历的个体和没有工作经历的个体的部分似然，然后将两个部分似然相乘。事实上，所有的 Cox 回归程序都有分层的选项。

表 4.4 的右边面板是将累犯数据通过工作经历分层后的结果。这个表格中最令人意外的结果是没有报告工作经历的结果。当你将一个变量分层后，你控制了这个变量但是你不会得到对这个变量的估计。通过分层控制一个变量的好处是你不用在意对那个变量的比例机会假设。但是这会对其他变量的估计造成影响吗？为了回答这个问题，分层方法的结果可以和表 4.1 的第 2 面板相比较。这些结果通过把工作经历放到模型里来控制这一变量，而不是通过分层的方法。在这

个例子中,两种结果之间的差异很小。但是在其他情况下,这种差异可能会比较明显。

分层方法看起来并不是一个特别有用的技术:它局限于分类变量,它不会对这些变量进行影响估计,并且它不会提供非比例性检验。但是,它依然有很多有用的地方,我们在第6章对重复事件的讨论中能看到一些。

第7节 | 时间尺度原点的选择

模型选择到目前都未被讨论的一个方面是时间始点的问题。尽管这一问题关乎参数模型和半参数模型,但是这一问题被推迟到现在讨论是因为比例机会模型提供了更为宽泛的选择可能性。在累犯的例子中,时间尺度的原点相对没有争议。释放日期是计算被释放后首次被捕的自然时间起点。相似地,如果研究者估计离婚机会的模型,那么结婚日期就是计算离婚机会的自然时间起点。

然而,在很多其他情况下,时间尺度的原点并非是现成的。即便在看起来很明确的例子中,仍然有商榷的余地。例如,在构造累犯事件的部分机会函数时,研究者可以令机会是个体的年龄或日历时间而非被释放的时间。年龄和日历时间同样是离婚机会可能的时间尺度。

既然利用部分似然法估计此类模型并没有困难,但是这么做合理吗?这依赖于研究实质性的考虑。如果已知机会对年龄有很强的依赖性而对其他起始时间

的依赖性较弱,那么年龄可能是定义时间尺度最合理的方式。或者,如果机会会随着以同样的方式影响所有样本成员的历史条件发生改变,那么日历时间也许是最好的尺度。

理论上,研究者是可以用公式来表示和估计比例机会模型的。在这一模型中,机会依赖于两个或以上的时间尺度。但是在实际操作中,这要求非常大的样本量或者图玛(Tuma, 1982)所总结的特殊条件。但是,即便这种方法是不可行的,研究者仍可以引入不同的时间尺度作为解释变量。例如,在估计一个离婚机会随结婚时间变化的比例机会模型时,研究者可以将日历年、丈夫的年龄和妻子的年龄作为解释变量。如果这些变量中有一些(对机会的对数)存在非线性的影响,有必要指定一个时变解释变量。如果这一影响(对风险的对数)是线性的,将开始的时间尺度作为一个解释变量就足够了,在这个例子中是指婚姻的开始时间。具体请参考埃里森(Allison, 2010: chapter 5)。

第8节 离散时间数据的Cox回归

部分似然方法的讨论假设时间是在连续的尺度上测量，因此两个事件不可能同时发生。在实际操作中，时间往往是按离散的单位测量的，尽管这一单位可能很小。许多数据集包括"连接"(ties)——两个或更多的个体表面上同时经历事件。因此，考虑到累犯的例子中，时间是以周为单位测量的，在很多周中有两个或更多的个体被逮捕。

为了处理有连接的数据，模型和估计方法都要做出改变。就模型来说，我们使用第2章介绍的离散时间模型，或者方程2.2的logit模型，或者方程2.4的补充log—log模型。如前所述，补充log—log模型的吸引力在于它暗含了连续时间上的Cox模型。

logit模型可以通过部分似然进行估计，但是计算要求(时间和存储)会很大，尤其是如果连接和个体的数量很多的情况下。对于补充log—log模型，合适的估计方法被称为边际似然法，它也需要很大的计算空间。在最近这几年，两种方法的算法都得到了很大的

发展,一些软件包(著名的有 SAS, Stata 和 R 软件里的 **coxph** 方程)都提供了进行分析的选项。但是这些方法对于有很多连接的数据集仍然不实用。

为了避免这些计算难题,许多近似的方法被提出。最广为接受的是布雷斯洛(Breslow, 1974)提出的方法。Breslow 方法是大多数 Cox 回归程序的默认选择,其使用了补充 log—log 模型进行边际似然的近似。如果没有连接的数据,Breslow 方法的公式会转而使用连续时间数据的部分似然方法。

然而,如果相对于处于风险中的个体的数量,发生在许多时间点的事件数量很大的话,比如说 15%或以上(Farewell & Prentice, 1980), Breslow 方法可能是不准确的。幸运的是,有一种更好的近似方法(Efron, 1977),很多软件包提供了这一选择,包括 SAS 和 Stata。Efron 方法的运行会比 Breslow 方法需要更多的时间,并且是默认选项(例如 R 软件里的 **survival** 程序)。当然,研究者也可以使用第 2 章里描述的离散时间极大似然方法。这些方法可以在不需要近似的情况下处理包含许多连接的大型数据集。

尽管累犯数据包含了一些连接,但是每周的被捕人数仅仅是处于风险中的人数的很小一部分。最大的比例(0.015 2)发生在第 50 周,在 330 个处于风险集里的人中有 5 个人被逮捕。因此,研究者采用部分似然法、边际似然法、Breslow 方法还是 Efron 方法都没有

什么区别。另一方面，对于第 2 章的升迁数据，在 10 个观察时间点中的 6 个，事件发生的数量在风险集的比例大于 0.15(参见表 2.1)，选择不同的处理连接的方法会产生明显不同的结果。

第9节 | 基于Cox模型的预测

许多人想要通过事件史模型来预测事件在未来的发生情况。如第3章所述,通过参数回归模型,这种预测很容易,尽管这种预测可能不是很可靠,尤其是当它们偏离观测数据的范围时。Cox回归模型预测的局限性很大,但是它们会避免过分偏离的危险。

研究者从多数Cox回归程序里得到的是一个预测生存函数,这个函数可以通过观测到的事件发生时间范围来评估。你可以得到关于你指定的解释变量的任何值的生存函数,或者是对于某些人的实际观测的数值,或者是完全假设的数值。例如,我们考虑表4.1的第1面板,这里没有任何时变解释变量。其中一个个体有以下的解释变量数值:接受资金援助、21岁、黑人、有工作经历、未婚、处于假释期,并且以往有过四次被指控的经历。表4.6显示了对这个人的预测生存函数。实际的生存函数有单独的一列,对应0到52周。表格仅仅报告了以5周为区间的结果,以及最后第52周的结果。

表 4.6 单个个体的预测生存函数

周	生存可能性	95%置信区间	
0	1.000 00	·	·
5	0.989 15	0.978 83	0.999 57
10	0.967 38	0.947 05	0.988 14
15	0.945 12	0.916 02	0.975 15
20	0.911 22	0.869 94	0.954 46
25	0.886 20	0.836 55	0.938 80
30	0.865 45	0.809 19	0.925 62
35	0.839 76	0.775 77	0.909 03
40	0.806 76	0.733 42	0.887 42
45	0.783 02	0.703 40	0.871 65
50	0.747 15	0.658 77	0.847 39
52	0.737 53	0.646 90	0.840 85

"生存可能性"这一列是对没有被逮捕的时间点估计的"生存"可能性。例如,第10周没有被逮捕,那么生存的可能性估计值是0.97,95%的置信区间是0.95到0.99。第40周没有被逮捕的生存可能性估计值是0.81,置信区间是0.73到0.89。使用这样一个表格,研究者可以进行其他很多种预测。例如,给定个体在第30周时依然是自由的情况下,在被释放之后的第30到第50周之间再次被捕的预测可能性是(0.865－0.747)/0.865＝0.136。

这个生存函数存在的问题是它会在第52周,也就是最后的观测周终止。我们对这之后会怎么样一无所知。如果生存可能性在任何一个星期低于0.50,那么我们可以估计一个被捕的中位值。如果生存可能性趋向于0的话,那么我们可以估计被捕的平均时间,就像

其他统计量一样。但是在这个例子中,我们只能对被释放之后的52周的范围里做出预测。

值得注意的是,生存函数往往不包含任何解释变量,而使用的是卡普兰和梅尔(Kaplan & Meier, 1958)发展的方法。Kaplan-Meier 方法被广泛应用于估计回归模型。但是,它假设每个人有相同的生存函数。对于检验两个或多个群组的生存函数是相同的零假设,我们有一些著名的方法(例如对数秩检验)。等价的假设也可以在 Cox 回归的框架下检验,仅仅包含一个分类解释变量就够了。不过 Cox 回归的美妙之处在于研究者同样可以控制其他解释变量。

第5章

多种类事件

在前面几章中,所有的事件被当作相同的。因此,在第2章中,我们没有区分不同种类的升职;在第3章和第4章中,所有的被捕事件也都被视为一样的。但事实常常不是如此。在有些情形下,将不同种类的事件混为一谈是完全不合适的。甚至即便这种处理是合适的,我们仍然需要单独检验不同类别的事件,从而进行更为精确的分析。

幸运的是,我们不需要新的方法。我们已经讨论过的单一事件的方法同样适用于多种类事件的分析。这些方法只不过要以更为复杂的方式应用其中。遗憾的是,围绕这一话题仍存争议,我认为这很大程度上归咎于存在很多种“多种类事件”这一事实。这一术语真实地描述了几种完全不同的情形,而每种都需要不同的方法进行分析。

第1节 | 多种类事件的分类

这一章的第一个尝试是要对不同的情形进行分类。简便起见，我们假设只有两类不同的事件。超过两类事件的推广是很容易的（一般地，事件种类的数目由分析者自己选择）。而且，我们继续假设事件是非重复的，把这一复杂的问题留到下一章讨论。

第一个主要类别——条件过程——表述如下：

1. 事件的发生或不发生是由一个因果过程决定的；在事件发生的前提下，第二个因果过程决定发生事件的种类。

很容易想到这一类别的例子。考虑一个买手机的事件，假设我们要在买苹果手机和安卓手机之间做出选择。要区分出导致两种事件的因果过程是不可能的。一个人首先决定要买一个手机，然后在这一前提下，决定要买的手机是苹果还是安卓手机。很可能是不同的解释变量影响了这两种决定。另一个例子是看

医生,我们将医生区分为骨科医生和牙科医生。我们通常认为拜访医生的决定和去拜访哪一种医生的决定是不同的。哪一决定在前并没有关系,重要的是两个决定是不同的。这些例子的共同之处在于,个体有一个目标(比如通讯),并且有达成这一目标的不同方法(苹果或安卓)。

对于这类“多种类事件”,合适的分析策略是和决定过程的结构是相同的。首先,研究者不考虑事件之间的差异,使用前面几章论述的事件史方法来构建事件发生的模型;然后,仅仅观测那些经历事件的个体,利用合适的技术来对决定事件类型的因果过程进行建模。常用的方法是二分 logit 分析(或者如果事件类别超过两种,则采用多分类 logit 分析)。

第二个“多种类事件”——平行过程——的广义类别符合以下描述:

2. 每种事件的发生都有不同的因果机制。

不同的因果机制意味着有不同的解释变量影响每种事件的发生,或者相同的解释变量有不同的系数或函数形式。与其给出宽泛的例子,我们不妨先将其细分为 4 个子类。在论述完 4 个子类之前,我们先暂缓对其分析方法的讨论。

2a. 一种事件的发生将个体从其他种类事件发生的风险里移除。

这又被称为“竞争性风险”，这类事件受到了生物统计学家和人口学家的广泛关注。经典的例子是竞争性原因导致的死亡。很明显，心脏疾病引发的死亡和癌症引发的死亡有着不同的因果过程。一个死于心脏疾病的人不再处于癌症引发死亡的风险里，反之亦然。这在社会科学中有许多其他的例子。例如，如果仅仅是因为不同的决策者经历了两种工作终止，那么很可能自愿的工作终止和非自愿的工作终止经历了不同的因果过程。一个人一旦离职了，那么他(她)就不可能处于被开除的风险中。并且一旦被炒了，一个人就不再拥有离职的选项。一个相似的例子是婚姻破裂，离婚和丧偶带来的婚姻终止显然是不同的。

2b. 一种事件的发生将个体从其他种类事件的观测中移除。

在人类迁移的研究中，研究者可能会区分国内移民和国际移民。由于移居国外而没有被继续观测的个体是很常见的。尽管这一个体不再被观测到，但是他(她)仍然存在国内移民的风险。

当然这个例子是非对称(asymmetric)的，因为国

内移民的个体仍然有国际移民的风险并且可以被继续观测。但我们很容易想象到对称(symmetric)的例子。例如,累犯的研究可能会区分暴力犯罪和非暴力犯罪。如果研究仅仅涉及被释放后的首次被捕经历,那么这一研究应被归到2b这一类别。

2c. 一种事件的发生既没有影响其他种类事件发生的风险,也没有影响对其观测。

尽管可能并不存在两类完全无关的事件,但是基于实际的考虑,在很多情形下仍然会将各事件视为无关的。例如,假设一个事件是选举时投票,另一个事件是离婚;或者也许一个事件是加薪,另一个事件是发生车祸。

2d. 一种事件的发生会提高或降低(但不会到0)其他种类事件发生的机会。

这类例子很常见。对于未婚女性,怀孕会提高结婚的机会。而婚姻反过来会提高生育的机会。职位提升会降低离职的机会。找到工作会降低被捕的风险。

第2节 | 平行过程的估计

现在我们考虑如何处理这四个子类。2c 的处理很简单。如果一个事件的发生对另一事件发生的风险或观测都没有影响的话，那么在对第二个事件的研究中，第一个事件可以被完全忽略。因此，这一类事件和我们在前面几章讨论的情形是一样的。

另一方面，如果事件的发生会提高或降低另一事件发生的机会（2d），那么第一个事件应当被考虑进对第二个事件的研究中。同样，我们也已经有做这类研究的方法。窍门是将第一类事件的发生定义为一个时变解释变量，将其放到对第二类事件的分析中。因此，在那个生物化学家的例子中，工作单位的声望这一时变变量被用以预测升职的发生。相似地，在累犯研究的例子中，一个虚拟变量被纳入模型之中以表明个体在每一周是否被雇佣。作为替代方法，研究者也可以创建一个测量找到工作后时间长度的变量。事实上，研究者完全可以将这两个变量作为解释变量。

2b 和前面已经讨论的右删截非常类似：在兴趣事

件发生之前,个体在某个时间点被移除出观测。这里的不同之处在于现在的删截即事件本身。抛开这一差异,最有效的分析策略其实是一样的。每一种事件类型都可以用前面几章所涉及的模型和方法单独分析。将个体移除出观测作为一种事件,它可以被视为个体在某一时间点被删截。例如,在分析国内迁徙的原因时,发生国际迁徙的个体(因此他们无法被追踪)可以被视为在发生国际迁徙的时间点被删截。特别值得注意的是,删截必须是随机的,而且我们要假定删截没有提供别的信息。

2a(竞争性风险)——一个事件的发生将个体从其他事件的风险中移除——是事件史研究中最常讨论的,因此,我们应当给予最多的关注。它和刚刚讨论到的 2b 非常相似,确实,它们的基本情况是相同的。应用于单一事件分析的方法可以对每一种事件进行单独分析。在分析每一事件时,个体被视为在其他事件发生时被删截。由于这一结果很重要,我们需要花一些时间在背景和存在的争论上。接下来我们看一个例子。

第3节　竞争性风险模型

有好几种方法可以解决竞争性风险的问题，最常见的是定义“特定类别”[type-specific，或“特定原因”(cause-specific)]的机会函数。假定存在 m 种不同类型的事件，令 $j=1, \cdots, m$ 作为区分不同事件类型的指标。给定个体在时间 t 上面临风险，令 $P_j(t, t+s)$ 为事件类型 j 在 t 与 $t+s$ 时间间隔之间发生的条件概率。注意，如果任何 m 个事件发生在时间 t 之前，那么个体在时间 t 上就不存在风险。

特定类别的机会比被定义为：

$$h_j(t)=\lim_{s\to 0}\frac{P_j(t, t+s)}{s} \qquad [5.1]$$

因此，每一事件类型都有其机会函数。据此，整体机会函数 $h(t)$，即任何 m 个事件发生的机会，等于所有类别机会函数之和。

对于每一个类别机会函数，研究者可以构建依赖时间和解释变量的模型。之前被讨论过的模型都可以作为候选。这些模型可能非常相似，或者对于每一种

事件而言,它们也可能完全不同。在任何情形下,可以看到数据的似然函数(即最大化的极大似然估计)可以作为因素被纳入到每种事件的单个似然函数。而且,这些因素看起来很像其他事件被删截时的单一事件的似然函数。因此,利用前面几章讨论的方法,研究者可以完成对单一事件的极大似然或部分似然估计。

在理论层面上,这一结果的重要性在于,竞争性风险模型并没有多少新的东西。而在实际层面上,单独估计每种事件的模型可以给予研究者在模型估计上极大的自由度和控制权。例如,研究者可以指定一种事件为 Weibull 回归模型而另一种事件为 Gompertz 回归模型。或者更可能的是,不同种类事件的模型可以有不同的解释变量,或者有相同的解释变量但却转换成了不同的形式。更重要的是,研究者可以忽略很少或不感兴趣的事件类型。例如,在一个婚姻破裂的研究中,如果研究者只关心离婚的原因,那么就没必要估计离婚和丧偶两种模型。

第 4 节 | 竞争性风险的实例

对于竞争性风险分析的例子，我们再次考虑累犯研究的数据(Rossi，Berk & Lenihan，1980)。这项以"对释放囚犯的过渡期援助"(TARP)闻名的研究是大规模的重复实验，这在第 3 章和第 4 章已经被讨论和分析。从得克萨斯和佐治亚监狱释放的接近 4 000 名犯人被随机指定了实验处理状态：包括多种经济援助和求职援助。在被释放后的一年内，这些人被追踪调查。并且在一年后，研究者对这一年中发生的任何逮捕的公开记录进行搜集。在这个例子中，分析被限定在 932 个佐治亚犯人。兴趣事件是释放后第一次被逮捕，因此，那些在一年时间内没有再次被捕的个体被处理为右删截。

依照涉嫌犯罪的类别，我们区分了两种不同类别的拘捕：财产犯罪(抢劫、入室盗窃、偷盗等)和其他犯罪。这一区分非常重要，因为我们预计经济援助能降低因财产犯罪而被逮捕的机会，经济因素正是刺激财产犯罪的元凶。与此相反，我们并不期待经济援助能

对非财产犯罪起很大作用。

由于被释放的确切日期是已知的,连续时间方法是最适合的。我们估计以下形式的 Cox 比例机会模型:

$$\log h_j(t) = a_j(t) + b_{j1}x_1 + b_{j2}x_2 + \cdots \qquad [5.2]$$

其中 j 下标表示有不同类别的系数以及对于每一种被捕类型,有不同的时间函数。解释变量包括资金援助、种族(白人或非白人)、教育(受教育年限)、被释放时候的婚姻状态、被释放时的年龄、性别、以往被捕的次数、财产类犯罪被指控的次数、表示被捕是否属于财产犯罪的虚拟变量和是否处于假释期的虚拟变量。由于所有的变量都不是时变变量,我们的分析大大简化。

我们首先估计一个不区分不同类型犯罪的模型。有 334 人经历了至少一次被捕,那么剩余的没有被捕的 598 人在 365 天里被删截。表 5.1 给出了估计的系数(前 3 列)。10 个自变量中有 6 个在 0.05 的水平上显著:被释放时的年龄、白人或非白人、以往财产犯罪被指控次数、因为财产犯罪被监禁、被捕次数和受教育程度。需要注意的是,资金援助的虚拟变量并没有显著的影响,甚至和我们预计的影响相反。因此,看起来资金援助并不是降低整体累犯率的有效措施。

表 5.1 不同犯罪类型的比例机会模型估计

解释变量	所有被捕			财产类被捕			非财产类被捕		
	b	z	$Exp(b)$	b	z	$Exp(b)$	b	z	$Exp(b)$
资金援助(D)[a]	0.121	1.06	1.128	0.201	1.34	1.222	0.007	0.041	1:007
释放时年龄	−0.034	−4.01**	0.966	−0.041	−3.38**	0.959	−0.027	−2.26*	0.973
白人(D)	−0.241	−2.03*	0.786	−0.353	−2.26*	0.703	−0.077	−0.42	0.926
男性(D)	0.501	1.62	1.650	0.111	0.32	1.117	1.387	1.94	4.005
已婚(D)	−0.222	−1.71	0.801	−0.332	−1.88	0.717	−0.073	−0.39	0.929
假释(D)	−0.211	−1.78	0.810	−0.147	−0.95	0.864	−0.302	−1.62	0.739
财产犯罪数	0.310	4.31**	1.364	0.318	3.27**	1.375	0.308	2.88**	1.361
因财产类犯罪坐牢(D)	0.424	3.09**	1.529	0.883	4.28**	2.417	−0.069	−0.36	0.933
被捕次数	0.018	3.81**	1.018	0.019	2.98**	1.019	0.016	2.35*	1.016
教育	−0.067	−2.67**	0.935	−0.053	−1.56	0.948	−0.083	−2.18*	0.921
对数似然		−2 172.9			−1 268.3			−892.7	

注：[a](D)表示虚拟变量。
* 在 0.05 水平上显著。
** 在 0.01 水平上显著。

但是，一种可能的情况是资金援助降低了财产类侵犯的机会而非其他类型的犯罪。其他变量对两种类型的犯罪也可能有不同的影响。为了检验这一可能性，我们将被捕事件划分为 197 个财产类被捕事件和 137 个非财产类被捕事件，然后分别对每种类型的被捕事件估计单独的比例机会模型。在估计财产类被捕事件模型时，首次被捕是因为非财产类案件的个体被处理为在被捕时间上删截。相似地，在对非财产类被捕事件的模型上，财产类被捕事件被处理为删截观测值。

在第一次被捕之后，个体仍然会被继续观测并且处于两种被捕的风险之中，这看起来像是人为制造的一个竞争性风险的例子。实际上，我们有理由认为这个例子应该被划分为 2d 类而不是 2a 类。无论如何，如果研究者坚信被释放之后首次被捕是以后重操犯罪事业的关键一步，那么我们可以仅仅关注那一次被捕经历。如果首次被捕是因为财产类侵犯，那么个体就不再处于非财产类侵犯的风险之中；反之亦然。表 5.1 给出了对两类被捕事件的估计结果。资金援助对两种类型的被捕经历都没有显著的影响。然而，它对财产类被捕的影响的偏离程度要比在总体估计中还要大。年龄、以往因财产犯罪被指控次数、以往被捕次数的影响对于两类事件几乎相同。另一方面，种族对财产类犯罪有显著的影响(白人的机会要比非白人低 30%)，但是对非财产类被捕并没有显著的影响。而且，以往

因为财产类犯罪入狱的经历显著提高了之后财产类犯罪的风险，但是对非财产类犯罪影响不大。最后，教育对非财产类犯罪有显著的负的影响，但是对财产类犯罪没有显著的影响。

那么对于这两类不同的事件，系数在统计上是否显著的不同？我们可以利用下面这个公式对每组系数进行差异检验：

$$z=\frac{b_1-b_2}{\sqrt{[s.e.(b_1)]^2+[s.e.(b_2)]^2}} \qquad [5.3]$$

其中 b_1 和 b_2 是要比较的两个系数，$s.e.(b_j)$是系数 b_j 的标准误。在无差异零假设下，z 统计量近似于正态分布。

下面是对于“财产类犯罪入狱”的结果(使用了表 5.1没有显示的标准误)：

$$2.70=\frac{0.883-(-0.069)}{\sqrt{0.296^2+0.192^2}}$$

显然，两个系数之间存在显著的不同。另一方面，对于财产类和非财产类犯罪来说，“白人”这个变量的系数之间没有显著的差异：

$$-1.15=\frac{-0.353-(-0.077)}{\sqrt{0.156^2+0.182^2}}$$

我们也可以检验两种类别犯罪的所有系数都相等的零假设。这可以通过计算总模型的对数似然，然后

减去两个犯罪类别的模型对数似然之和来检验。将这个差值乘以−2就是似然比卡方检验：

$$23.8=-2(-2\,172.9-(-1\,268.3-892.7))$$

自由度是要检验的系数的数量，在这个例子中是10，p值是0.008。因此我们拒绝无差异零假设，认为至少有一对系数存在显著差异。

然后，我们看到区分不同类别的事件会带来不同的关于解释变量影响的结论。相似地，没有区分不同类别的事件可能会带来误导性的结果。但是需要注意的是，研究者不要把事件类别分得太细，因为这样会使很少的事件数量分成一个或多个事件类型。在表5.1中，3个模型的样本量是一样的，但是“所有被捕事件”面板的标准误(没有显示)要比两种类别的事件分别估计的标准误小。这是由于标准误(和统计效力)更多地依赖于事件的数量而不是总体的样本量。

第5节 | 不同种类事件间的依赖

在刚刚提到的竞争性风险的方法中，我们对每类事件估计的模型将所有其他类型的事件处理为删截。正如第2章所定义的，这显然属于随机删截，因为我们没有控制其他事件发生的时间。如果删截是随机的话，那么我们假设这是没有信息的。在累犯的例子中，我们必须假设，个人因财产类犯罪而被捕这一事实没有告诉我们关于非财产类被捕机会的任何信息。相似地，如果一个人因为非财产事件而被捕，那么它也没有告诉我们这个人因财产类事件而被捕的风险是高还是低。

如第2章所述，我们没办法验证非信息性假设。我们能做的是那一章所介绍的敏感性检验，即检验多种极端的情形。另一件需要牢记的事情是，如果回归包含了许多影响所有事件类型的解释变量，那么删截包含信息的可能性更低。实际的问题是在调整解释变量之后，删截是否就会变得有信息了。

第6节 | 累计发生函数

有另一种竞争性风险的方法。它是基于累计发生函数的,而且不需要删截是非信息性的(Marubini & Valsecchi, 1995; Fine & Gray, 1999)。这种相对比较新的方法在一些统计软件包里已经可以找到,而且有时候它也是处理竞争性风险优先选择的方法。这种方法在以预测为主要目的的应用里很有用。但是,我还是认同平蒂利(Pintilie, 2006)的观点,即这个方法并不适合做因果推断。

累计发生函数的最初动机在于,当存在竞争性风险的时候,通过标准方法(例如 Kaplan-Meier 或者 Cox 回归)估计的生存函数有一个麻烦的特点。具体来说,如果你使用这些函数去估计个体在特定的时间点处于不同状态的概率,这些概率相加可能大于1。我们来看累犯数据。使用 Kaplan-Meier 方法,在一年的观察期的结束,我估计个体因为财产犯罪而被逮捕的可能性是0.229,因非财产犯罪被逮捕的可能性是0.168,不被逮捕的可能性是0.642。遗憾的是,这三个概率之和达

到了 1.039。尽管这看起来不是很大的问题,但是在模型预测时可能影响很大。

解决方案是使用累计发生函数而不是生存函数。对于事件类型 j 的累计发生函数是:

$$CI_j(t) = \Pr(T < t,\ J = j) \qquad [5.4]$$

换句话说,它简单地表示了事件类型 j 在时间 t 之前发生的概率。这可以通过在没有任何假设的情况下进行直接地估计。SAS 有一个内置的宏命令叫作 **cumincid**,可以用来生成这些估计,而可下载的宏命令 **cif** 有更多的功能。Stata 有一个用户贡献的命令叫作 **stcompet** (Coviello & Boggess, 2004)。

图 5.1 给出了财产类犯罪和非财产类犯罪的累计发生曲线,这是通过 **stcompet** 命令分析的。它显示的是对于两种类型的事件,被捕事件作为自释放之后的

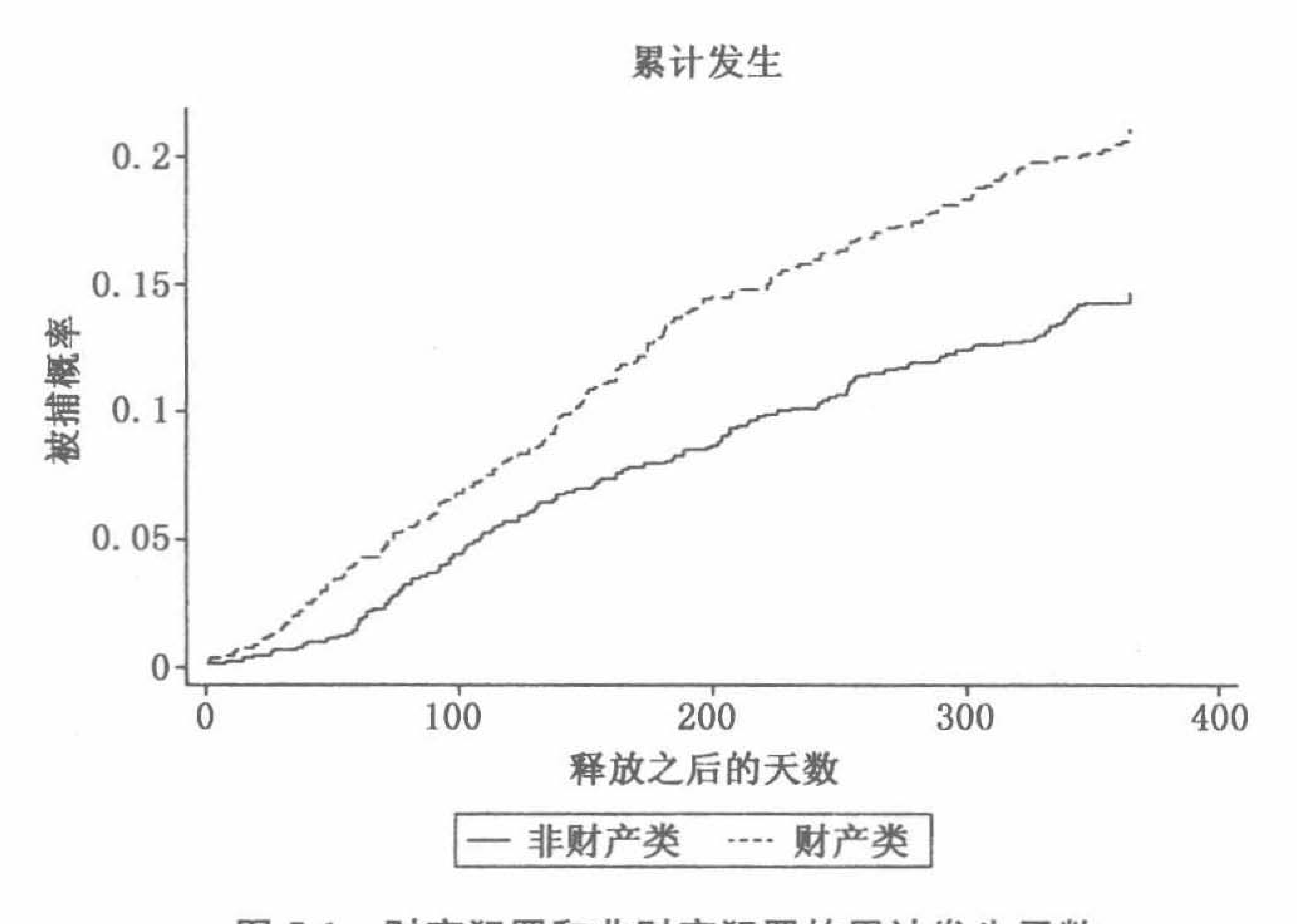


图 5.1　财产犯罪和非财产犯罪的累计发生函数

表 5.2 不同犯罪类型的子分布机会模型估计

解释变量	所有被捕			财产类被捕			非财产类被捕		
	b	z	$Exp(b)$	b	z	$Exp(b)$	b	z	$Exp(b)$
资金援助(D)[a]	0.121	1.06	1.128	0.236	1.55	1.266	0.004	0.02	1.004
释放时年龄	−0.034	−4.01**	0.966	−0.040	−2.96**	0.960 6	−0.022	−1.93	0.978
白人(D)	−0.241	−2.03*	0.786	−0.351	−2.17*	0.704	−0.010	−0.06	0.990
男性(D)	0.501	1.62	1.650	0.043	0.11	1.043	1.436	2.01*	4.204
已婚(D)	−0.222	−1.71	0.801	−0.308	−1.75	0.735	−0.035	−0.17	0.966
假释(D)	−0.211	−1.78	0.810	−0.108	−0.70	0.898	−0.279	−1.56	0.756
财产犯罪数	0.310	4.31**	1.364	0.275	2.70**	1.317	0.258	2.52*	1.294
因财产类犯罪坐牢(D)	0.424	3.09**	1.529	0.887	4.26**	2.429	−0.184	−0.96	0.832
被捕次数	0.018	3.81**	1.018	0.015	2.16*	1.015	0.011	1.67	1.011
教育	−0.067	−2.67**	0.935	−0.049	−1.31	0.952	−0.075	−2.03*	0.928
对数似然		−2 172.9			−1 288.2			−914.7	

注：[a](D)表示虚拟变量。

* 在 0.05 水平上显著。

** 在 0.01 水平上显著。

天数的函数的估计概率。例如，在200天里，因为财产犯罪被逮捕的累计概率是0.145；而对于非财产犯罪，预计概率是0.087。

在一个通过部分似然方法估计的比例"子分布"机会模型里，这一方法可以将解释变量加入进去(Fine & Gray, 1999)。为了应用这个方法，Stata有一个内置的命令叫作 **stcrreg**。对于SAS，有一个用户编写的宏命令叫作 **pshreg**(Kohl & Heinze, 2012)。

表5.2显示的是对于财产类犯罪和非财产类犯罪的子分布机会模型的估计结果，这是通过 **stcrreg** 命令得到的。第1面板(对于所有被捕的事件)等价于表5.1的结果，因为这里没有竞争性风险。另外两面板的结果和表5.1也没有明显不同，尽管许多 z 值很小。就

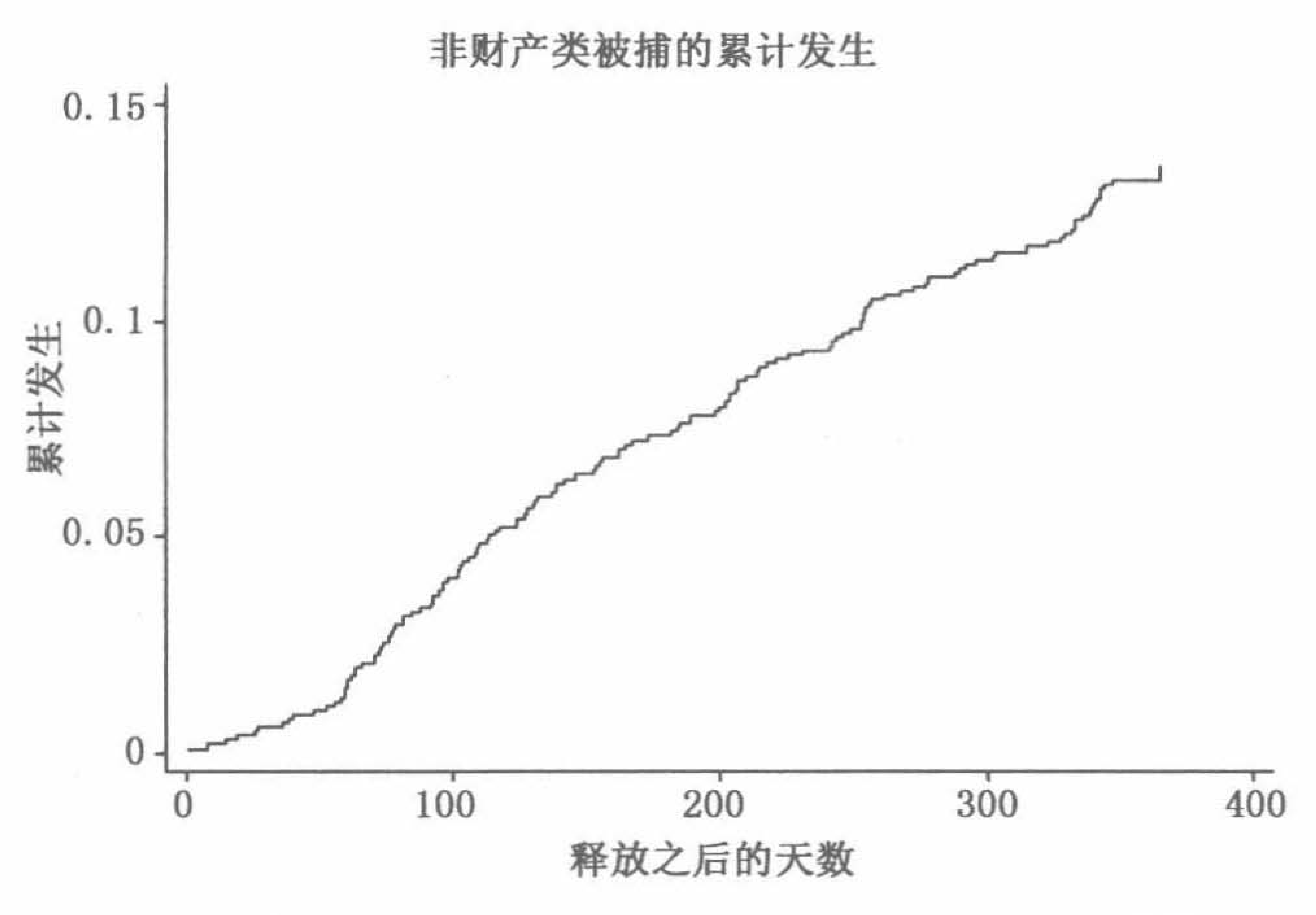


图5.2 在所有解释变量上都是均值的个体的非财产犯罪的预测累计发生函数

像我前面所说的，我更倾向于选择表 5.1 中的比例机会模型进行因果推断，主要是因为子分布模型的估计值整合了一些概念上不同的现象。无论如何，子分布估计值在进行预测时更加有效。例如，图 5.2(由 Stata 里的 **stcurve** 命令生成)显示了一个假想的在所有解释变量上都是均值的个体，其非财产犯罪的预测累计发生函数。像这样的图像可以在任何指定的解释变量取值上获得。

第6章

重复事件

社会科学家研究的大部分事件是可重复的(反复出现的),而且很多事件史数据集包含了每个个体的重复事件,例如工作变动、生育、结婚、离婚、逮捕、犯罪和就医。在本书的第一版中,我观察到"仅仅有少量的文章是关于重复事件的分析的"(例如,Gail, Santner & Brown, 1980; Prentice, Williams & Peterson, 1981; Tuma, Hannan & Groeneveld, 1979; Flinn & Heckman, 1982a, 1982b)。但是自从第一版以来的近30年间,关于重复事件的文献已经大大丰富。对于这些文献很好的总结可以参见阿伦、博根和耶辛(Aalen, Borgan & Gjessing, 2010),库克和劳利斯(Cook & Lawless, 2007),胡加德(Hougaard, 2000),或者纳尔逊(Nelson, 2003)。

重复事件的分析有很多方法,我们很难在这短短的一章里对这些方法一一介绍。但是,我会试着解释和总结一些主要的取向,并且将关注点放到研究者如何从中选择更加合适的取向。

我们首先从扩展第 5 章的实例开始。样本包含了 932 个从佐治亚州立监狱释放之后被观察一年的个体。在前面的分析中，分析的重点是在被释放之后的第一次被捕。然而，如表 6.1 所示，在一年的跟踪观察里，有些人被捕不止一次。准确地说，132 个人被捕超过一次，他们贡献了除了上一章的分析之外的 195 次事件。忽略这些信息是一种损失。为了简化进一步分析，我们假设重复的被捕事件都是单一种类。也就是说，我们不去区分财产类犯罪和非财产类犯罪。

表 6.1　被捕次数的频次分布

被捕次数	人　数
0	598
1	202
2	88
3	28
4	10
5	4
6	2

第1节 | 重复事件的计数分析

重复事件最简单的分析取向是忽略事件的发生时间,而仅仅关注发生在个体身上的事件数量。如果(1)你没有任何时变变量,或者(2)你想假设解释变量在整个观察期内有相同的影响,那么这是最好的分析方法。在这些条件下,关于事件发生时间的数据并没有多少有用的信息。

对于计数数据最合适和可靠的模型是负二项回归模型。泊松回归模型是负二项回归的一个特例,但是它往往和数据的拟合度很差,这是由于“过度散布”(overdispersion)的问题(Allison, 2012)。设 Y_i 是个体 i 经历的事件次数。令 Y_i 有一个期望值是 λ_i 的负二项分布。回归模型可以表示为:

$$\log \lambda_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} \quad [6.1]$$

也就是说,事件的期望数量是一系列解释变量的对数线性函数。对方程 6.1 中的系数 b 的解释和 Cox 比例参数模型的系数的解释几乎等价。这是因为机会本身就是事件的期望数值。

表 6.2 重复被捕的回归模型

解释变量	负二项计数模型			Cox 回归,间隔时间				Cox 回归,共享异质		
	b	z	$Exp(b)$	b	z	稳健 z	$Exp(b)$	b	z	$Exp(b)$
资金援助(D)[a]	0.148	1.36	1.160	0.137	1.52	1.33	1.147	0.142	1.30	1.153
释放时年龄	−0.033	−4.12**	0.968	−0.030	−4.53**	−3.64**	0.970	−0.033	−4.14**	0.967
白人(D)	−0.153	−1.36	0.858	−0.160	−1.71	−1.40	0.852	−0.156	−1.38	0.855
男性(D)	0.372	1.39	1.450	0.344	1.50	1.19	1.411	0.372	1.38	1.451
已婚(D)	−0.052	−0.44	0.949	−0.078	−0.79	−0.64	0.925	−0.069	−0.57	0.934
假释(D)	−0.323	−2.77**	0.724	−0.304	−3.15**	−2.86**	0.738	−0.324	−2.78**	0.723
财产犯罪指控次数	0.278	3.68**	1.321	0.256	4.58**	3.88**	1.292	0.284	3.74**	1.328
因财产类犯罪坐牢(D)	0.343	2.64**	1.409	0.316	2.90**	2.41*	1.372	0.335	2.58**	1.398
被捕次数	0.013	2.71**	1.013	0.012	3.36**	2.65**	1.012	0.013	2.77**	1.014
教育	−0.060	−2.52*	0.942	−0.054	−2.76**	−2.13*	0.947	−0.059	−2.51*	0.942

注:[a](D)表示虚拟变量。
* 在 0.05 水平上显著。
** 在 0.01 水平上显著。

表 6.2 的第 1 面板呈现的是使用第 5 章里的解释变量的负二项回归的结果(估计值可以通过 Stata 的 **nbreg** 命令和 SAS 的 **genmod** 程序过程获得)。总体而言,这些结果和表 5.1 里的"所有被捕"面板非常相似,但是有一些例外。"白人"变量在这里有一个更小的系数并且不显著,但是在表 5.1 里是显著的。而"假释"变量在这里有一个更大和显著的系数,但是在表 5.1 并不显著。

在这个例子里,所有的个体都被追踪观察了一年。然而,负二项回归模型可以很容易被扩展到允许不同个体的不同观察时期。这可以通过定义一个作为观察时期长度对数的变量来实现,并且把这个变量纳入到模型中作为一种"抵消变量"。这里的抵消变量是指系数限定到 1.0 的变量。

第2节 | 基于间隔时间的方法

如果存在时变解释变量,或者研究者感兴趣的是解释变量的影响是否随时间发生变化,那么负二项回归模型就不令人满意了。因此,无论哪种情况,我们需要一个允许每个个体有多个观察记录的更加复杂的取向。为了达成这一目的,我们以事件发生为标志将整个观察期分为不同的区间,每个区间生成单独的记录:从观察开始到第一个事件的发生是一个记录,相邻两个事件之间构成一个记录,最后一个事件和观察期结束之间的区间构成一个记录。对于 932 个在被捕数据集中的个体,总共有 1 447 个这样的区间。

我们将这些区间视为不同的观察对象,首先分析“间隔时间”,即每个区间开始时间和结束时间之间的差。本质上说,我们在每次事件发生之时,将时钟重设到 0 开始计算。表 6.2 的第 2 面板给出了基于时间间隔的渐变 Cox 回归估计的结果,这些通过 Stata 的 **stcox** 命令和 SAS 的 **phreg** 程序来完成。

我们注意到 Cox 的结果和负二项回归的结果非常

相似,尤其是系数和暗含的机会比。但是 Cox 回归的 z 值总要比负二项回归模型对应的 z 值稍大。这是由于 Cox 模型的标准误总是更小,也是因为软件并不知道每个个体贡献了样本中的多个观察值。所有的时间间隔都被视为相互独立的,实际上,经常被逮捕的人的时间间隔很小,而很少被逮捕的人的时间间隔很长。这是“聚类”(clustered)观察值的典型情况,我们需要修正统计依赖性以做出有效的推断。

最简单的解决方法是利用休伯(Huber, 1967)和怀特(White, 1980)提出的“三明治”方法来计算稳健标准误。在 Stata 中,这可以简单地将 **cluster(id)** 放到 Cox 回归命令里执行,其中 **id** 指的是每个个体所具有的独特编号。在表 6.2 中,“稳健 z 值”这一列报告了使用稳健标准误计算的 z 值。修正之后,z 统计量在一定程度上变小了,并且和负二项回归得到的结果大致相同。

对于很多应用来说,修正标准误是唯一需要做的事情。但是标准 Cox 回归模型的系数也不是完美的,这主要有两个原因。第一,如果每个个体的多个观测值之间存在相互依赖,标准 Cox 回归估计值在统计上就不是有效的。也就是说,他们会比所需要的拥有更多的样本异质性。第二,有一些系数会被削弱,趋向于 0,但是这可以通过其他方法来修正。

一种方法是包括“共享异质性”这一部分的 Cox 回

归。基本的模型就是一个将随机截距项加入到标准Cox模型的随机效应或者混合效应模型：

$$\log h_{ij}(t)=a(t)+b_1x_1+b_2x_2+e_i \quad [6.2]$$

在这个方程中，$h_{ij}(t)$是第 i 个个体发生第 j 个事件的机会，t 是自从上次事件发生以来的时间长度。误差项 e_i 是共享异质性。它可以被视为所有未被测量并且不随时间变化的变量对个体 i 的机会的影响。误差项有时候又被视为未被观察到的异质性。我们假设误差项 e_i 是一个以 0 为均值、以 θ 为固定方差的随机变量，并且它在统计上独立于 x 变量。θ 越大，每个个体的多个观测值之间的相互依赖越强。我们还必须假设 e_i 有一个特定的概率分布，例如正态(SAS)或对数 gamma(Stata)。

共享异质模型可以通过轮廓似然函数进行估计。轮廓似然函数一般化了 Cox 的部分似然估计(Therneau & Grambsch，2000)。表 6.2 的最后一面板显示了结果(使用 Stata 程序)。整体而言，共享异质模型的结果和报告稳健标准误的简便 Cox 回归的结果很相似。共享异质模型报告的系数在一定程度上更大一些，起码有些系数在统计上显著了。这符合我们的预期，但是这些差异并不明显。在 Stata 的结果中，θ(e_i的方差)的估计值是 0.759，并且当 $\theta=0$ 时，检验高度显著。但是当相同的模型使用 SAS 估计时，θ 的估计就小了很多，并且在统计上不再显著。这种差异可

能是由于两种软件不同的误差分布设定导致的。

需要注意:不论有没有稳健标准误,包含共享异质的 Cox 回归模型的估计要比简便 Cox 回归需要更长的运算时间。在我的笔记本电脑上,表 6.2 的前两面板的 Cox 回归在 Stata 里仅仅各需要 1 秒钟的运算时间。但是第 3 面板的共享异质模型需要长达 15 分钟的运算时间。如果你要分析指数异质模型(就像第 3 章里讨论的那些),运算就会更加可控。通过使用 Stata 里的 **streg** 命令,我将一个 Weibull 模型(Cox 模型的特殊形式)和共享异质以及与表 6.2 中相同的解释变量相拟合。这一运算仅仅需要 3 秒,而且得到的结果和共享异质 Cox 回归的结果非常相似。

现在我们有了回答那些负二项回归无法解决的问题的工具。例如,被捕的机会会随着上一次被捕增加或是降低?表 6.3 给出了拟合时间间隔的 Weibull 模型的结果。这一模型加入了一个额外的预测变量:区间的顺序变量。例如,第一个区间的变量值是 1,第二个区间的变量值是 2,以此类推(我之所以使用 Weibull 模型是因为我已经厌倦了共享异质 Cox 模型运行的等待时间)。在第 1 面板,我们看到的是包含稳健标准误的简便 Weibull 模型更正了重复观察之间的依赖性。对于顺序数字,我们看到了一个正的系数,并且有着所有变量中最大的 z 值。在每次被捕之后,再次被捕的机会上升约 40%。

表 6.3 重复被捕的 Weibull 和 Cox 回归模型

解释变量	Weibull,稳健 z			Weibull,共享异质			Cox 回归,起点时间		
	b	z	$Exp(b)$	b	z	$Exp(b)$	b	z	$Exp(b)$
顺序号	0.334	7.43**	1.396	−0.010	−0.11	0.990			
资金援助(D)[a]	0.135	1.48	1.145	0.142	1.27	1.153	0.140	1.30	1.150
释放时年龄	−0.029	−3.77**	0.972	−0.034	−4.11**	0.966	−0.031	−3.68	0.969
白人(D)	−0.119	−1.17	0.888	−0.158	−1.36	0.854	−0.166	−1.40	0.847
男性(D)	0.340	1.28	1.404	0.375	1.37	1.455	0.355	1.18	1.426
已婚(D)	−0.061	−0.56	0.941	−0.066	−0.54	0.936	−0.086	−0.67	0.918
假释(D)	−0.270	−2.75**	0.763	0.330	−2.74**	0.719	−0.308	−2.79	0.735
财产类犯罪指控次数	0.224	3.95**	1.251	0.292	3.66**	1.339	0.270	3.92	1.309
因财产类犯罪坐牢(D)	0.296	2.48*	1.344	0.338	2.54*	1.402	0.326	2.40	1.385
被捕次数	0.009	2.48*	1.009	0.013 9	2.66*	1.014	0.012	2.65	1.012
教育	−0.055	−2.52*	0.946	−0.061	−2.51*	0.941	−0.057	−2.13	0.945

注：[a](D)表示虚拟变量。

* 在 0.05 水平上显著。

** 在 0.01 水平上显著。

但这种影响很可能是人为操作的。因为所有的个体被追踪观察了一年,那么唯一取得更大顺序号的方法是经历更多次被捕,因为这样会使区间时间变得更短。而更短的区间意味着更高的被捕机会。

我们可以通过一个共享异质模型来控制这一问题,表 6.3 的第 2 面板显示了结果。共享异质 Weibull 模型允许个体之间存在未被观察到的差异的可能性,这些差异很可能导致有些人有更高的被捕机会(并且因此被捕多次),而有些人的机会比较低(被捕次数很少)。在这个模型里,顺序号码的系数很小并且 z 值是最低的。因此,通过控制未被观察到的异质性,我们排除了很可能被解释为每次被捕事件对下次被捕的因果影响。

第3节｜基于起点时间的方法

在基于间隔的方法中，在每次事件发生的时候，时钟被重设为0，因此允许机会依赖自上次事件发生以来的时间长度。这在大多数应用中可能是最好的方法。但有时我们允许机会依赖于其他时间标尺可能更有意义。例如，在被捕的例子中，我们怀疑被捕的机会仅仅依赖于个体从监狱里被释放出来的时间，而不是上次被捕以来的时间。

这个想法很容易实现，在方程6.2里，只需要将时间变量定义为从起点（从监狱被释放出来）以来的时间，而不是从最近一次事件以来的时间。为了估计这个模型，我们能继续使用之前的数据集，也就是以事件为分界，将每个记录对应于一个时间区间。但是现在，我们不再将区间的时间长度作为因变量，而是指定区间的起始时间和结束时间，它们都从共同的起点（在这个例子中是从监狱被释放）开始算起。大多数Cox回归软件包（也包括一些指数模型的软件）都允许设定每次观测的起始和结束时间。

在被捕数据集中，有一个变量 *begin*，这是每个区间从释放开始算起的起始时间；另一个变量 *end*，是那个区间从释放开始算起的结束时间。例如，数据集中第二个人有两次被捕事件，一次在第 114 天，另一次在第 153 天。所以第一个记录是从 0 到 114，第二个记录是从 114 到 153，第三个（右删截）区间是从 153 到 365。

使用这个形式的数据，我利用表 6.3 的解释变量估计了一个 Cox 模型，排除了顺序数字（因为我们的共享异质模型已经将它排除出了预测变量）。因为我们有重复事件，我选择了稳健标准误以调整个体内的事件集群。表 6.3 的最后一面板显示了结果。最合适的比较是和表 6.2 的第 2 面板，它报告了基于间隔时间而不是起点时间的 Cox 回归结果。很显然，结果非常相似，就像前面的情形一样。

指定起点时间的一个吸引力在于，它允许我们检验一个或多个变量的影响是否随着整个观察期而变化。我们可以通过制定一个预测变量和自起点以来时间的交互项来检验这个问题。例如，我们估计下面形式的模型：

$$\log h_{ij}(t) = a(t) + b_1 x_1 + b_2 x_2 + b_3 x_2 t$$

其中 t 表示从起点开始算起的时间。这个模型告诉我们的是 x_2 对机会的影响是自起点开始算起的时间的线性函数。这和我们在第 4 章（方程 4.5）为了检验比

例机会假设所做的分析方程类似。

我对累犯数据进行了分析，把起点以来的时间（月）和之前被捕的次数生成交互项代入。交互项的影响是负的并且在统计上显著（$p=0.003$），这意味着之前被捕次数的影响随着被释放时间而下降。具体来说，在被释放的时间上，之前的被捕次数每增加1次，被捕机会增加2.8%。但是每多1个月，这一影响下降0.3%。这意味着在被释放6个月以后，之前的被捕次数每增加1次对机会的影响仅仅是1%。在第12个月的时候，这一影响基本为0。

第4节 | 扩展

我们之前讨论的重复事件的方法可以有多种扩展。例如,第5章竞争性风险方法可以很容易地应用于重复事件数据。在累犯的例子中,假设我们区分非财产和财产类犯罪。然后,为了顾及财产类犯罪的模型,我们简单地将所有以非财产类犯罪结束的区间处理为在结束的时间删截的区间。相似地,非财产类犯罪的模型将所有以财产类犯罪结束的区间处理为在结束的时间删截的区间。时变解释变量可以简单地使用第4章介绍的方法处理,即使用区间分割方法或程序陈述方法。

重复事件可以很容易通过使用第2章介绍的离散时间极大似然法进行分析。我们将个人的事件史处理为一系列的个人数据记录,每个记录是一个离散时间点,代表从观察开始的时间到删截或者事件发生的时间。但是,对于事件发生之后或者删截之后的时间并没有记录,因为个体不再处于风险之中。

如果事件是重复的,我们对每个个体被观测的时

间点创建一个离散时间记录，这个过程不考虑事件是否已经发生。当只有一类事件时，如果事件在一个特定的时间点发生，那么因变量被记为1，否则是0。如前所述，我们只需要利用 logistic 回归（或者补充 log—log 回归）就可以预测结果。

不同之处在于，我们现在必须要考虑每个个体的多个观测之间的依赖性。幸运的是，处理这种依赖性的方法已经被发展出来，甚至是连续时间方法。我所描述的数据结构和二分因变量的面板数据是一样的。现在已经有很多方法用于处理这类数据：稳健标准误、一般估计方程（GEE）、随机效应（混合）模型和固定效应模型（Allison，2005，2009）。

对于离散时间数据的这些取向，间隔时间模型和起点时间模型之间的早先差别越来越小。对于间隔时间模型，你只需要将自上次事件发生以来的时间长度作为因变量包括在内。而对于起点时间模型，自起点以来的时间长度是一个解释变量。同样你也可以将两个都包括进去。



结　论

不管研究的是哪一领域,研究者都需要注意,有多少现象可以被概念化为事件或者事件的顺序。事件无处不在,可以是从微观层次上的微笑,到宏观层次上的战争或者生物灭绝。

这本书重点关注回归的方法,研究事件发生的概率或机会如何依赖于解释变量。就像其他回归方法一样,事件史/生存分析既可以用来检验解释变量对事件发生的影响,也可以对未来事件的发生进行预测。

尽管我们讨论了几种不同的取向,但是它们有几个共同的吸引力:

- 事件发生的准确时间会增加因果解释的准确性以及减少不确定性;
- 处理不同类型的删截,尤其是右删截的能力;
- 将时变解释变量(有些方法不支持)纳入模型之中的能力。

在这些方法之中,Cox 回归最为流行。其所基于的比例机会模型比一些更为普遍的参数模型需要的假

设更弱。它可以处理连续时间和离散时间数据。简便和有效的程序在标准的统计软件里都可以找到。这个方法可以将时变解释变量考虑进去。基于这些原因，Cox 回归是大多数情形下的第一选择。

但是 Cox 回归也有局限性。不像第 3 章讨论的参数模型，Cox 回归不能处理左删截或者说区间删截的一般形式。而且这一模型很难用于预测，因为这个模型仅仅能预测相对机会而非绝对机会。尽管我们可以基于 Cox 模型修正预测生存函数，但那些功能由于删截的问题被严重减弱了。

第 2 章讨论的离散时间方法也提供了一种替代 Cox 回归的方法。由于可以通过 logistic 回归执行，它们将生存分析放到了一个简单易懂的框架里进行讨论。考虑到机会对时间的依赖，这一方法有很大的灵活性，它允许多种时间标尺并且允许检验时间依赖性的假设。最后，不像 Cox 回归，在数据集很大以及时间高度分散(有很多连接的事件时间)的情况下，离散时间模型很容易运行。

附　录

极大似然

极大似然的原理是选择一组数值作为参数估计，这些数值能够最大化实际观测到的数据的似然性(可能性)。第一步是将数据的似然性表达为一组未知参数的函数。这个附录将解释当部分数据是右删截的时候，如何构建参数回归模型。

让我们假设有一个 n 个独立个体的样本 ($i = 1, \cdots, n$)。对于每个个体，数据包含(t_i, d_i, x_i)，其中 t_i 是事件发生的时间或删截的时间，d_i 是一组虚拟变量(如果 t_i 未被删截，数值为1，否则为0)，x_i 是一列解释变量(包括作为截距的1)。如果观测是独立的，整体样本的似然性仅仅是单个个体观测似然性的结果，即：

$$L = \prod_{i=1}^{n} L_i \qquad [A1]$$

对于未删截的观测，$L_i = f_i(t_i)$，其中 f_i 是个体 i 的密度函数。注意，f_i 的下标表明密度依赖于解释变量，所以因个体而异。对于删截数据，$L_i = S_i(t_i)$，其中

S_i 是生存函数。因此，$S_i(t_i)$是对于个体 i，事件发生在 t_i 之后的概率。我们可以把这些方程结合为：

$$L=\prod_{i=1}^{n} f_i(t_i)^{d_i} S_i(t_i)^{1-d_i} \qquad [A2]$$

在这个方程里，d_i 充当了一个转换器，打开了未被删截的观察值的密度，也打开了右删截个案的生存函数。对于删截没有信息性的参数模型，这个方程仍然成立。

本书中最简单的模型是指数(常数机会)模型。设 λ_i 是个体 i 的机会，那么指数模型有以下密度函数：

$$f_i(t)=\lambda_i e^{-\lambda_i t}$$

以及生存函数是：

$$S_i(t)=e^{-\lambda_i t}$$

把这些表达式代入方程 A2 并作一些代数转换，我们就能得到：

$$L=\prod_{i=1}^{n} \lambda_i^{d_i} e^{-\lambda_i t} \qquad [A3]$$

因为最大化函数的对数等价于最大化函数自身，为了简便，我们进行对数转换，这可以将乘数转化为和以及幂转化为系数：

$$\log L=\sum_{i=1}^{n} d_i \log \lambda_i-\sum_{i=1}^{n} \lambda_i t_i \qquad [A4]$$

此时，我们利用假设 $\log \lambda_i=\mathbf{\beta x}_i$，其中 $\mathbf{\beta}$ 是一个行向量，结果有：

$$\log L = \boldsymbol{\beta}\sum_{i=1}^{n} d_i \mathbf{x}_i - \sum_{i=1}^{n} t_i \exp(\boldsymbol{\beta}\mathbf{x}_i) \qquad [A5]$$

现在我们成功地将似然表达为未知参数 $\boldsymbol{\beta}$ 的方程。下一步是利用一些数值运算(一般是迭代法)去找到能最大化 $\log L$ 的 $\boldsymbol{\beta}$ 的值。牛顿—拉夫森算法(Newton-Raphson algorithm)常常能满足这一目的,并且能够呈现估计值的标准误。更详细的内容,参见 Kalbfleisch 和 Prentice (2002)。

部分似然

部分似然和极大似然很类似，第一步是构建依赖于未知参数和观测数据的似然函数。第二步是寻找最大化函数的参数值。然而，一般的似然方程是样本中所有个体似然性的结果。与之不同，部分似然是被观测到发生的所有事件似然性的结果。因此，

$$PL = \prod_{K=1}^{K} L_K \qquad [A6]$$

其中 PL 是部分似然，K 是样本中事件的总数。

为了理解 L_K 如何构建，我们考虑表 A1 中的假设案例。现在我们有 10 个个案的样本，但只观测到 5 次事件，其余 5 个个案被删截。3 个观测在时间 12 处被删截，我们假定这是因为研究在此时结束。观测 4 在时间 5 处被删截，观测 6 在时间 9 处被删截。在这两个个案中，删截的发生可能是因为死亡，或者因为有意不再参与研究，或者无法在之后的后续访谈中找到个体。

为了简便，观测按照时间 t_i 排列，t_i 代表删截或事件发生的时间。第一次事件在时间 2 处发生在个体 1

身上。此时,所有 10 个个体都面临事件发生的风险。我们现在要问:给定事件在时间 2 处发生,那么它发生在个体 1 身上而非其他 9 个个体之一身上的概率是多少? 这个概率记为 L_1。它可以被表示为:

$$L_1 = \frac{h_1(2)}{h_1(2) + h_2(2) + \cdots + h_{10}(2)} \qquad [A7]$$

表 A1 部分似然估计的计算例子

i	t_i	k	L_k
1	2	1	$e^{\mathbf{bx}_1}/(e^{\mathbf{bx}_1}+e^{\mathbf{bx}_2}+\cdots+e^{\mathbf{bx}_{10}})$
2	4	2	$e^{\mathbf{bx}_2}/(e^{\mathbf{bx}_2}+e^{\mathbf{bx}_3}+\cdots+e^{\mathbf{bx}_{10}})$
3	5	3	$e^{\mathbf{bx}_3}/(e^{\mathbf{bx}_3}+e^{\mathbf{bx}_4}+\cdots+e^{\mathbf{bx}_{10}})$
4	5*		
5	6	4	$e^{\mathbf{bx}_5}/(e^{\mathbf{bx}_5}+e^{\mathbf{bx}_6}+\cdots+e^{\mathbf{bx}_{10}})$
6	9*		
7	11	5	$e^{\mathbf{bx}_7}/(e^{\mathbf{bx}_7}+e^{\mathbf{bx}_8}+\cdots+e^{\mathbf{bx}_{10}})$
8	12*		
9	12*		
10	12*		

注:i 表示个体;t_i 表示事件发生的时间或个体 i 的删截时间;k 表示事件。

* 删截。

和前面一样,在此 $h_i(t)$表示个体 i 在时间 t 处的机会。因此,我们赋予在时间 2 处经历事件的个体以机会,令其除以在时间 2 处面临风险的所有个体的机会值之和。尽管方程 A7 直观上非常有吸引力,但是正式的推导实际上非常冗长(Tuma, 1982),这里就不再给出过程。

不考虑存在时间和解释变量依赖性的模型选择问

题，L_i 的表达式始终成立。但是，在比例机会模型下，它可以得到简化。对于模型：

$$h(t)=\exp[a(t)+\mathbf{bx}_i]=\exp[a(t)]\exp[\mathbf{bx}_i] \quad [A8]$$

其中 $\mathbf{x}_i$ 是对于个体 i 的解释变量的列向量，$\mathbf{b}$ 是系数的行向量。将其带入 L_i 的表达式，每一项消除 $\exp[a(t)]$，得到：

$$L_1=\frac{\exp[\mathbf{bx}_1]}{\exp[\mathbf{bx}_1]+\exp[\mathbf{bx}_2]+\cdots+\exp[\mathbf{bx}_{10}]} \quad [A9]$$

完全不考虑非指定函数 $a(t)$，正是这次消减使得估计系数向量 $\mathbf{b}$ 成为可能。

L_2 通过同样的方式构建。给定事件在时间 4 处发生，L_2 是事件发生在于时间 4 面临风险的个体 2 而非其他人身上的概率。唯一的区别在于个体 1 已经经历了一次事件，在时间 4 处不再面临风险。因此，

$$L_1=\frac{\exp[\mathbf{bx}_2]}{\exp[\mathbf{bx}_2]+\exp[\mathbf{bx}_3]+\cdots+\exp[\mathbf{bx}_{10}]} \quad [A10]$$

表 A1 给出了方程 L_3，L_4 和 L_5。

注意，每一个 L_k 的数值并不依赖于第 k 次事件发生的准确时间。它可能发生在第 $(k-1)$ 次事件之后和第 $(k+1)$ 次事件之前的任何时间点，并且大小相

同。仅仅是时间的顺序影响部分似然。

一旦部分似然被构建,就可以利用牛顿—拉夫森算法,像对一般似然函数一样将其最大化(Kalbfleisch & Prentice, 2002; Lawless, 2002)。

非重复事件的离散时间似然函数

在第 2 章，我提出如果事件是不重复的，那么对于每个个体有多个记录的 logistic 回归模型的估计并不需要修正依赖性。尽管对依赖性的担心是理所应当的，但是它并不适用于此。在这个例子中，多个观测的创建并不是一个特殊的方法；相反，它遵循了对数据的似然函数进行因子化的方法(Allison, 1982)。

下面是简要的讨论。假设我们有一个 n 个个案的样本，其中前 r 个个案未被删截，而剩下的 $n-r$ 个个案被删截。按照原有的数据形式，每个个体有一个观测，那么数据的似然函数可以被写为对于 n 个观测的概率的乘积：

$$L=\prod_{i=1}^{r}\Pr(T_i=t_i)\prod_{j=r+1}^{n}\Pr(T_i>t_i) \qquad [A11]$$

其中 T_i 是表示事件发生时间的随机变量，t_i 是个体 i 的具体时间，要么是被观察到的事件发生时间，要么是删截时间。方程 A11 中的每一个概率都可以通过以下方式因子化。如果一个未删截的个案有 $t_i=5$，那么我

们有：

$$\Pr(T_i = 5) = P_{i5}(1 - P_{i4})(1 - P_{i3})(1 - P_{i2})(1 - P_{i1}) \quad [A12]$$

其中 P_{it} 是离散时间机会，即给定事件还没有发生，在时间 t 时，事件发生的条件概率。这个因式分解遵循了条件概率的定义。方程 A12 中五个项式中的每一个都表现出它来自一个完全不同的、相互独立的二项观测值。

相似地，如果一个观测在时间 4 的时候被删截，它对似然值的贡献可以写为：

$$\Pr(T_i > 4) = (1 - P_{i4})(1 - P_{i3})(1 - P_{i2})(1 - P_{i1}) \quad [A13]$$

再次，这个表达式和四个独立的二项观测的似然值等价。

参考文献

Aalen, O., Borgan., O., & Gjessing, H. (2010). *Survival and event history analysis: A process point of view*. New York, NY: Springer.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. In S. Leinhardt (Ed.), *Sociological methodology*, pp. 61–98. San Francisco, CA. Jossey-Bass.

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. Cary, NC: SAS Institute.

Allison, P. D. (2009). *Fixed effects regression models*. Thousand Oaks, CA: Sage.

Allison, P. D. (2010). *Survival analysis using SAS: A practical guide* (2nd. ed.). Cary, NC: SAS Institute.

Allison, P. D. (2012). *Logistic regression using SAS: Theory and application* (2nd. Ed.). Cary, NC: SAS Institute.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics*, *30*, 89–99.

Brown, C. C. (1975). On the use of indicator variables for studying the time-dependence of parameters in a response time model. *Biometrics*, *31*, 863–872.

Cook, R. J., & Lawless, J. F. (2007). *The statistical analysis of recurrent events*. New York, NY: Springer.

Coviello, V., & Boggess, M. (2004). Cumulative incidence estimation in the presence of competing risks. *Stata Journal*, *4*(2),103–112.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, Series B 34, 187–202.

D'Agostino, R. B., Lee, M.-L., Belanger, A. J., Cupples, L. A., Anderson, K., & Kannel, W. B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham heart study. *Statistics in Medicine*, *9*, 1501–1515.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, *72*, 557–565.

Farewell, V. T., & Prentice, R. L. (1980). The approximation of partial likelihood with emphasis on case-control studies. *Biometrika*, *67*, 273–278.

Fine, J. P., & Gray, R. J. (1999). Proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, *94*, 496–509.

Flinn, C. J., & Heckman, J. J. (1982a). New methods for analyzing individual event histories. In S. Leinhardt (Ed.), *Sociological methodology*, pp. 99–140. San Francisco, CA: Jossey-Bass.

Flinn, C. J., & Heckman, J. J. (1982b). Models for the analysis of labor force dynamics. In G. Rhodes & R. Basmann (Eds.), *Advances in econometrics*., pp. 35–95. New Haven, CT: JAI.

Gail, M. H., Santner, T. J., & Brown, C. C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics*, *36*, 255–266.

Glasser, M. (1967). Exponential survival with covariance. *Journal of the American Statistical Association*, *62*, 561–568.

Heckman, J. J., & Singer, B. (1982). The identification problem in econometric models for duration data. In W. Hildebrand (Ed.), *Advances in econometrics*. Cambridge, UK: Cambridge University Press.

Holford, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics*, *36*, 299–305.

Hougaard, P. (2000). *Analysis of multivariate survival data*. New York, NY: Springer.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 221–223.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). New York, NY: Wiley.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457–481.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data* (2nd ed.). New York, NY: Springer.

Kohl, M., & Heinze, G. (2012, August). *PSHREG: A SAS macro for proportional and nonproportional substribution hazards regression with competing risk data*. Technical report. Available at http://cemsiis.meduniwien.ac.at/fileadmin/msi_akim/CeMSIIS/KB/programme/tr08_2012-PSHREG.pdf

Laird, N., & Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association, 76*, 231–240.

Lawless, J. F. (2002). *Statistical models and methods for lifetime data* (2nd. ed.). New York, NY: John Wiley.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Long, J. S., Allison, P. D., & McGinnis, R. (1979). Entrance into the academic career. *American Sociological Review, 44*, 816–830.

Mantel, N., & Hankey, B. (1978). A logistic regression analysis of response-time data where the hazard function is time-dependent. *Communications in statistics—Theory and methods*, A7, 333–347.

Marubini, E., & Valsecchi, M. G. (1995). *Analysing survival data from clinical trials and observational studies*. New York, NY: Wiley.

Nelson, W. B. (2003). *Recurrent events data analysis for product repairs, disease recurrences, and other applications*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Peterson, A. V. Jr. (1976). Bounds for a joint distribution with fixed sub-distribution functions: application to competing risks. *Proceedings of the National Academy of Sciences, 73*, 11–13.

Pintilie, M. (2006), *Competing risks: A practical perspective*. New York, NY: Wiley.

Prentice, R. L, & Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer. *Biometrics,* 34,: 57–67.

Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case control *Studies. Biometrika, 66*, 403–411.

Prentice, R. L., Williams, B. J., & Peterson, A. V. (1981). On the regression analysis of multivariate failure data. *Biometrika, 68*, 373–374.

Preston, S., Heuveline, P., & Guillot, M. (2000). *Demography: Measuring and modeling population processes*. New York, NY: Wiley-Blackwell.

Rossi, P. H., Berk, R. A., & Lenihan, K. J. (1980). *Money, work and crime: Some experimental results*. New York, NY: Academic.

Singer, B., & Spilerman, S. (1976). The representation of social processes by Markov models. *American Journal of Sociology, 82,* 1–54.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika, 69*, 239–241.

Sørensen, A. B. (1977). Estimating rates from retrospective questions. In D. R. Heise (Ed.), *Sociological methodology*. San Francisco, CA: Jossey-Bass.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival 'data: Extending the Cox model*. New York, NY: Springer.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, *72*, 20–22.
Tuma, N. B. (1976). Rewards, resources and the rate of mobility: A nonstationary multivariate stochastic model. *American Sociological Review*, *41*, 338–360.
Tuma, N. B. (1982). Nonparametric and partially parametric approaches to event history analysis. In S. Leinhardt (Ed.), *Sociological Methodology*, pp. 1–60. San Francisco, CA: Jossey-Bass.
Tuma, N. B., & Hannan, M. T. (1978). Approaches to the censoring problem in analysis of event histories. In K. F. Schuessler (Ed.), *Sociological methodology*. San Francisco, CA: Jossey-Bass.
Tuma, N. B., Hannan, M. T., & Groeneveld, L. D. (1979). Dynamic analysis of event histories. *American Journal of Sociology*, *84*, 820–854.
White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838.
Zippin, C., & Armitage, P. (1966). Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. *Biometrics*, *22*, 665–672.

译名对照表

accelerated failure time models	加速失效时间模型
asymmetric	非对称的
cause-specific	特定原因的
censoring	删截
competing risks	竞争性风险
conditional density	条件密度
continuous time	连续时间
discrete time	离散时间
episode-splitting method	区间分割法
event	事件
fixed censoring	固定删截
frailty model	异质模型
hazard	机会
hazard rate	机会率
heterogeneity	异质性
intercept	截距
interval	间隔
left censoring	左删截
life table	生命表
likelihood function	似然函数
likelihood ratio chi-square test	似然比卡方检验
log-linear model	对数线性模型
longitudinal data	纵贯数据
Markov process	马尔科夫过程
maximum likelihood estimation	极大似然估计
Newton-Raphson algorithm	牛顿—拉夫森算法
partial likelihood function	部分似然函数
person-year data	人—年数据
program statement method	程序陈述法
proportional hazard model	比例机会模型
repeated event	重复事件

right censoring	右删截
risk	风险
risk set	风险集
single sample	单样本
symmetric	对称的
tie	连接
time-variant variable	时变变量
time-constant variable	不随时间变动的变量
transition rates	转换率
type-specific	特定类别的

格致方法・定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用 logistic 回归分析（第二版）
14. logit 与 probit：次序模型和多类别模型
15. 定序因变量的 logistic 回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析（第二版）
26. 纵贯研究（第二版）
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析（第二版）
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图解代数：用系统方法进行数学建模
37. 项目功能差异（第二版）
38. Logistic 回归入门
39. 解释概率模型：Logit、Probit 以及其他广义线性模型
40. 抽样调查方法简介
41. 计算机辅助访问
42. 协方差结构模型：LISREL 导论
43. 非参数回归：平滑散点图
44. 广义线性模型：一种统一的方法
45. Logistic 回归中的交互效应
46. 应用回归导论
47. 档案数据处理：生活经历研究
48. 创新扩散模型
49. 数据分析概论
50. 最大似然估计法：逻辑与实践
51. 指数随机图模型导论
52. 对数线性模型的关联图和多重图
53. 非递归模型：内生性、互反关系与反馈环路
54. 潜类别尺度分析
55. 合并时间序列分析
56. 自助法：一种统计推断的非参数估计法
57. 评分加总量表构建导论
58. 分析制图与地理数据库
59. 应用人口学概论：数据来源与估计技术
60. 多元广义线性模型
61. 时间序列分析：回归技术（第二版）
62. 事件史和生存分析（第二版）